Four Essays in Multiple Testing and Prediction

Dissertation

for the Faculty of Economics,
Business Administration and Information Technology of the
University of Zurich

to achieve the title of
Doctor of Philosophy
in Economics

presented by

Dan Christian Wunderli
from Meilen, Zurich

approved in April 2012 at the request of
Prof. Michael Wolf, Ph.D.
Prof. Dr. Ashok Kaul

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorizes the printing of this Doctoral Thesis, without thereby giving any opinion on the views contained therein.

Zurich, 4.4.2012

Chairman of the Doctoral Committee: Prof. Dr. Dieter Pfaff

Contents

Part I

Overview ("Rahmenpapier" as required by the Faculty)

1.1 Multiple Testing

The underlying problem of all essays is the following, which can be solved by multiple testing methods. Controlling individual errors of type one with respect to one null hypothesis $H_{0,s}$, such as the well-known α of an individual test, does not ensure that multiple errors of type one with respect to a family of null hypotheses $\{H_{0,s}\}_{s=1}^{p}$ are controlled. That is, carrying out $s = 1, \ldots, p$ individual tests of $H_{0,s}$ at level α ensures that the probability of falsely rejecting each single $H_{0,s}$, one s at a time, is no larger than α. However, in testing p null hypotheses, one usually cares about the probability of falsely rejecting one or more null hypotheses within the family $\{H_{0,s}\}_{s=1}^{p}$. This is the familywise error rate FWE that is defined as follows

$$\text{Familywise error rate FWE} = P(\text{Number of falsely rejected null hypotheses} \geq 1) \tag{1.1}$$

where $P(\cdot)$ denotes the probability mechanism.

Carrying out p individual tests at individual significance level α does not ensure that the probability of falsely rejecting one or more null hypotheses (FWE) is no larger than α. As I show in a realistic simulation setup in Table 2.2 in Essay 2, this may mean that the probability of falsely labeling a nonexistent treatment effect as statistically significant is around 40% by individual t-testing. Furthermore, in an empirical context, Essay 2 illustrates to which extent the probability of falsely rejecting one or more null hypothesis (FWE) is ignored by individual t-testing: the FWE can be as high as 90% in individually t-testing six treatment effects at individual level $\alpha = 5\%$.

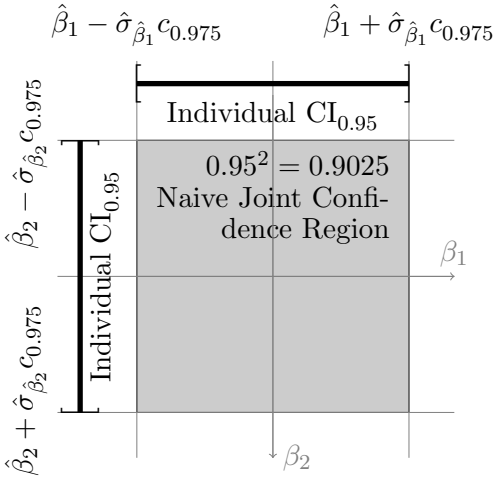


Figure 1.1: Product of two individual confidence intervals $\text{CI}_{1-\alpha}$ has joint coverage $(1-\alpha)^2$ for independent $\hat{\beta}_1$, $\hat{\beta}_2$

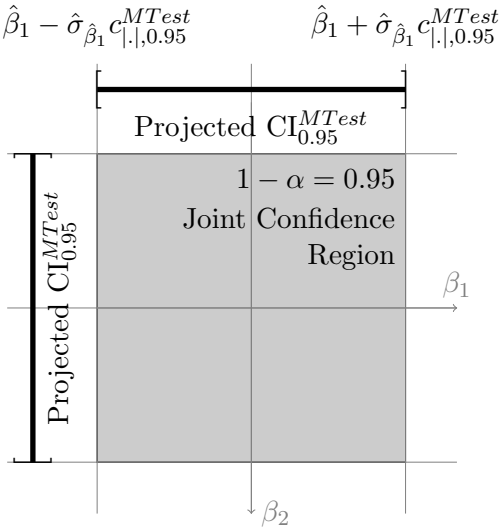


Figure 1.2: Projected confidence intervals $\text{CI}^{MTest}_{0.95}$ from 95% joint confidence region are larger than individual confidence intervals $\text{CI}_{0.95}$ in Figure 1.1[1]

To facilitate the comprehension of the problem mentioned above, consider two confidence intervals $\text{CI}_{1-\alpha}(\beta_1)$ and $\text{CI}_{1-\alpha}(\beta_2)$ for two independent estimators $\hat{\beta}_1, \hat{\beta}_2$. These two individual confidence intervals are depicted as two bracketed lines in Figure 1.1.

An essential insight of multiple testing is that the naive rectangular region spanned by these two individual confidence intervals, $\text{CI}_{1-\alpha}(\beta_1) \times \text{CI}_{1-\alpha}(\beta_2)$, only has coverage $(1-\alpha)^2$, not $1-\alpha$. Thus, the product of two such individual $1-\alpha$ confidence intervals results in an $(1-\alpha)^2$ confidence region, whose joint coverage is smaller than $1-\alpha$ [1]; the $\alpha = 5\%$ case is shown in Figure 1.1.

To construct a joint confidence region for p parameters at level $1-\alpha$, instead of the naive $(1-\alpha)^p$ region, one can generally apply the method of Romano and Wolf (2005). If the $\hat{\beta}_1, \hat{\beta}_2$ as above are correlated, the $1-\alpha$ joint confidence region is still larger than the naive joint confidence region, except for perfectly correlated $\hat{\beta}_1, \hat{\beta}_2$. Hence, the edges of the $1-\alpha$ joint confidence region are larger than the individual confidence intervals at individual level $1-\alpha$, as illustrated by comparing the vertical β_2 axes of Figure 1.1 and Figure 1.2.

Section 1.1 of Essay 1 and Section 2.1 of Essay 2 elaborate on multiple testing in the context of the corresponding essay.

1.2 Single-Step versus Step-Wise Multiple Testing

The multiple testing method in Romano and Wolf (2005) as such is a step-wise multiple testing method. That is, if the goal is testing, then step-wise multiple testing to increase the statistical power of the testing procedure makes sense. In that sense, the emphasis of step-wise multiple testing is on rejecting as many false null hypotheses as possible within the family $\{H_{0,s}\}_{s=1}^{p}$. This is the case in Essay 1 and in Essay 2. Furthermore, control of multiple error rates like the false discovery proportion FDP[2] only makes sense if the notion of a true null hypothesis exists. In case of unknown population parameters, for which a true value β_s exists, inferring on whether $H_{0,s} : \beta_s = 0$ is true or false makes sense.

If the goal is setting up a joint confidence region[3], however, a single-step procedure needs to be used. The reason is the following, as illustrated in Figure 1.3. Suppose that the first null hypothesis $H_{0,1} : \beta_1 = 0$ is rejected in the first step by using the multiple critical value c_1, while the second $H_{0,2} : \beta_2 = 0$ is only rejected in the second step by using the multiple critical value $c_2 < c_1$. By using c_1 in the β_1 dimension and using the smaller c_2 in the β_2 dimension, the resulting rectangular region does no longer have joint coverage $1-\alpha$. The same holds for using c_1 in the β_1 dimension and using c_2 in the β_2 dimension.

[1] Except for perfectly correlated $\hat{\beta}_1, \hat{\beta}_2$

[2] Or its expectation, the False Discovery Rate FDR

[3] Or a joint prediction region as in Essay 3

Nonetheless, by controlling the k-FWE

$$k\text{-FWE} = P(\text{Number of falsely rejected null hypotheses} \geq k) \tag{1.2}$$

with a single-step procedure, generalized joint confidence regions (GJCR) result, as depicted in Figure 1.2. In Essay 3, generalized joint prediction regions are introduced by controlling the k-FWE. If the FWE ($k = 1$) is controlled, rectangular joint confidence regions[4] result.

Last but not least, it is worth mentioning here that using elliptical regions to construct rectangular regions is not optimal; the resulting rectangular joint confidence regions are too large, as indicated in Figure 1.4. For the Scheffé approach, which is based on elliptical regions as well, other problems ensue as shown in Section 3.3.3 of Essay 3 and in Proposition 4.3.2 of Essay 4.

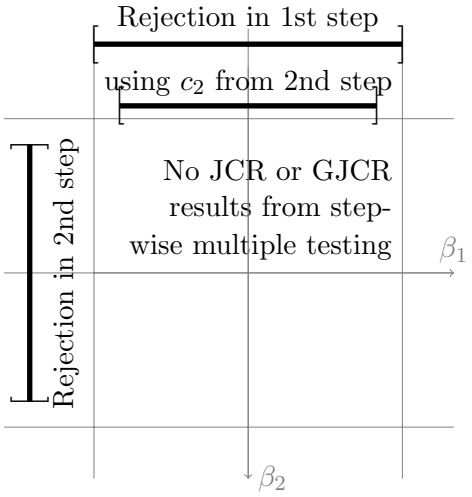


Figure 1.3: Using critical values from stepwise multiple testing does not result in (generalized) joint confidence region

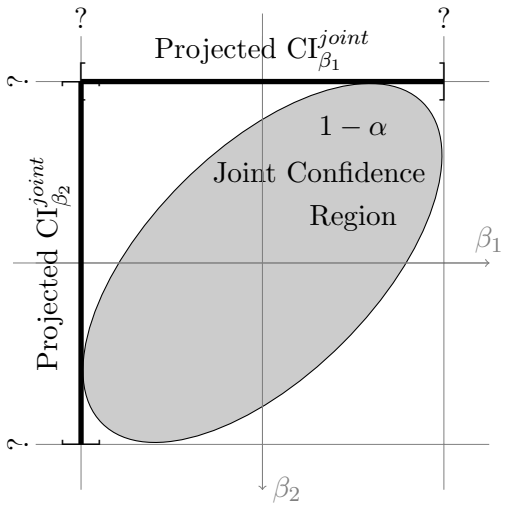


Figure 1.4: Individual CI's inferred from ellipsoidal joint region as in a test of one joint null hypothesis, e.g. F-test.

1.3 Contributions of the Four Essays

The main contribution of Essays 1 to 3 is to highlight that ignoring multiple error types can be a big mistake in a variety of contexts. Unfortunately, many researchers do not seem to be aware of this fact. The contribution of Essay 4 is that I explain and mathematically prove why so-called Scheffé bands are flawed for the construction of joint prediction regions or joint confidence regions.

On a technical level, not every multiple error type one makes sense in each application. In Essay 1 and in Essay 2, the notion of true null hypotheses exists, which is not the case for

[4]Or rectangular joint prediction regions

predictions in Essay 3. This determines what kinds of multiple error types one can be used, as briefly explained in Section 1.2 above.

In Essay 1, fund of funds (FoF) managers need to identify outperforming hedge funds out of a large universe of candidate funds. We study if such a selection can be successfully carried out by only looking at the track records of the available funds, using step-wise multiple testing methods that control the False Discovery Proportion. In particular, at a given point in time, we determine which funds significantly outperform a given benchmark at the same time as, crucially, accounting for the fact that a large number of funds are examined at the same time. Then, the equal-weighted or the global minimum variance portfolio of the outperforming funds is held for one year, after which the selection process is repeated. When backtesting this strategy on two particular hedge fund universes, we find that the resulting FoF portfolios have attractive return properties compared to the $1/N$ portfolio (that is, simply equal-weighting all the available funds) but also when compared to two investable hedge fund indices. This essay has been published as a chapter within the Oxford Handbook of Quantitative Asset Management.

In Essay 2, I expose the risk of false discoveries in the context of multiple treatment effects. A false discovery is a nonexistent effect that is falsely labeled as statistically significant by its individual t-value. Labeling nonexistent effects as statistically significant has wide-ranging academic and policy-related implications, like costly false conclusions from policy evaluations. I reexamine an empirical labor market model by using multiple testing methods and I provide simulation evidence. By merely using individual t-values at conventional significance levels, the risk of labeling probably nonexistent treatment effects as statistically significant is unacceptably high. Individual t-values even label a number of treatment effects as significant, whereas multiple testing indicates false discoveries in these cases. Tests of a joint null hypothesis such as the well-known F-test control the risk of false discoveries only to a limited extent and do not optimally allow for rejecting individual hypotheses. Multiple testing methods control the risk of false discoveries in general while allowing for individual decisions in the sense of rejecting individual hypotheses. This essay is under review at the journal Quantitative Economics of the Econometric Society.

In Essay 3, we show how to construct bootstrap joint prediction regions (JPRs) for predictions $[\hat{y}_{t+1}, \ldots, \hat{y}_{t+H}]$. These JPRs miss the path of H realized future values $[y_{t+1}, \ldots, y_{t+H}]$ at k or more prediction horizons with probability $\alpha_{k\text{-FWE}}$ at most in the following sense. The probability that the realized future path $[y_{t+1}, \ldots, y_{t+H}]$ lies outside our k-FWE joint prediction region at k or more of the H prediction horizons is no larger than $\alpha_{k\text{-FWE}}$, at least asymptotically. Previous approaches to construct such forecast bands that control the FWE ($k = 1$) are either multiple testing flawed, as is the case for Scheffé bands in Jordà and Marcellino (2010)[5], or merely of heuristic nature, as is the case for the so-called neighboring paths method of Staszewska-Bystrova (2010)[6]. In addition to the proofs of asymptotic validity, we

[5]These multiple testing deficient properties are proved in Essay 4

[6]There are no proofs of asymptotic validity; there is only simulation evidence for some VAR models.

provide simulation evidence that shows the superior finite-sample properties of our JPRs. We equip researchers and practitioners with a flexible tool to accurately quantify the uncertainty associated with prediction paths. This essay is a working paper version.

Essay 4 proves why Scheffé bands are flawed in light of basic insights from multiple testing. This essay is meant as a potential part or appendix of Essay 3, because it is not customary to publish a separate paper on why someone else's method is problematic. I mathematically prove why Scheffé bands have a number of deficient properties. These deficiencies partly show up in the two papers that introduce Scheffé bands, namely Jordà (2009) and Jordà and Marcellino (2010), as well as in the numerical results of Essay 3. This essay is a working paper version as well.

1.4 Other research during my Ph.D. studies

On a meta-level, an interesting question arising from Essay 1 is how one can identify outperforming weighting strategies out of a potentially infinitely large pool of candidate strategies. This question spurred me on to do some research on the verge of multiple testing and optimization with my colleague Gregor Reich from the University of Basel. We did a presentation of some results at the Institute of Computational Economics 2011 at the University of Chicago. Unfortunately, the results weren't concrete enough then to get proper attention, so we put the project on hold.

Furthermore, I contributed as a coauthor to the article "A Test of the Extreme Value Type I Assumption in the Bus Engine Replacement Model" with Bradley J. Larsen (Massachusetts Institute of Technology), Florian Oswald (University College London), and Gregor Reich (University of Basel). This article developed out of a course work at the Institute on Computational Economics 2011 at the University of Chicago; it is currently under review at the journal Economics Letters of Elsevier Publishing. The article is concerned with a common distribution assumption on the unobservable error term within dynamic discrete choice models, which is a dynamic programming problem. Although this article is concerned with statistics as well as optimization, it seemed to be too far off the topic of multiple testing to include it within this Ph.D. thesis.

Part II

The Four Essays

Essay 1

Fund-of-Funds Construction

Fund-of-Funds Construction by Statistical Multiple Testing Methods

Michael Wolf[1]
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

Dan Wunderli[2]
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
dan.wunderli@econ.uzh.ch

March 2010

Abstract

Fund-of-funds (FoF) managers face the task of selecting a (relatively) small number of hedge funds from a large universe of candidate funds. We analyse whether such a selection can be successfully achieved by looking at the track records of the available funds alone, using advanced statistical techniques. In particular, at a given point in time, we determine which funds significantly outperform a given benchmark while, crucially, accounting for the fact that a large number of funds are examined at the same time. This is achieved by employing so-called multiple testing methods. Then, the equal-weighted or the global minimum variance portfolio of the outperforming funds is held for one year, after which the selection process is repeated. When backtesting this strategy on two particular hedge fund universes, we find that the resulting FoF portfolios have attractive return properties compared to the $1/N$ portfolio (that is, simply equal-weighting all the available funds) but also when compared to two investable hedge fund indices.

KEY WORDS: Bootstrap, familywise error rate, fund-of-funds, performance evaluation.

JEL CLASSIFICATION NOS: C12, C14, C22, G11.

Note: This essay has been published as a chapter within The Oxford Handbook of Quantitative Asset Management; Scherer, B., and Winston, K. (Ed.), Oxford University Press, 2014

[1]Research supported by the Swiss National Science Foundation (NCCR FINRISK, Module A3).

[2]Many thanks to Ashok Kaul, Andrea Heuson, Iwan Meier, and Chayawat Ornthalanai for helpful comments. I would also like to thank Eurekahedge Inc. for the kind support, the Financial Management Association for organizing the FMA European Conference 2008 in Prague, and the LSE for organizing the LSE Alternative Investments Research Conference. I gratefully acknowledge financial support from the Swiss Banking Institute.

1.1 The Challenge

A fund-of-funds (FoF) manager or an institutional investor faces the challenge of selecting a (relatively) small number of 'good' hedge funds from a large universe of candidate funds. We shall address the problem of fund selection from a statistical point of view. The analysis will be based solely on the track records of the individual managers. Arguably, the track record constitutes the single most important piece of information to judge the quality of a fund manager.[3] But making sense of the track records is a non-trivial task.

If we want to answer the question whether a particular fund manager is skilled based on his track record, we can use a statistical test. Such a test declares a fund manager skilled if his alpha with respect to a suitable benchmark is statistically proven to be positive 'beyond a reasonable doubt', say a doubt threshold of 5%. This doubt threshold, say 5%, is denoted by the *significance level* of the test. By design, there is only a small chance then, say 5%, that a lucky manager passes the test, that is, gets wrongly identified as skilled.[4] Importantly, this logic assumes that *only one* manager is tested. If *many* managers are tested at the same time, the small individual doubts accumulate to a large global doubt. In other words, it now becomes very likely that some lucky managers will pass the test. This is undesirable for investment purposes. In general, only skilled managers will continue to outperform, while lucky managers will not.

The following analogy might help illustrate this dilemma. Imagine a person claims to have — some, though not necessarily perfect — extrasensory perception (ESP). A possible test consists of secretly tossing a coin ten times and having the person predict the outcome of each toss. It would then be reasonable to identify the person as possessing ESP if she scores at least nine correct predictions. The logic is that somebody guessing completely at random has a chance of about 1.1% to score at least nine correct predictions. As a result, there is only a small chance that an 'ignorant' person passes the test by chance.[5] But now consider 1,000 persons taking the test at the same time (perhaps because we put out a related job ad) and assume they are all ignorant. One would expect $0.011 \times 1000 = 11$ persons to pass the test by chance alone, that is, to get lucky. And the probability that at least one person will pass the test by chance alone, if they all guess independently of each other, is $1 - (1 - 0.011)^{1000} = 99.998\%$.

If our goal is to select the skilled managers from a large universe of candidates, we face a similar challenge as Cinderella:

[3] To be sure, there may be other pieces of information as well, such as the general background of the manager, his investment philosophy, the size and location of his office, etc. However, such factors are not easily quantifiable and/or available and so they will be left out for the statistical analysis.

[4] Imposing a significance level of 0% is not possible, as it would imply that no manager, based on a finite track record, could ever be found skilled, no matter how impressive his track record may be.

[5] Again, if we did not allow for a small chance of an ignorant person passing the test, based on a finite number of tosses, nobody could ever be declared as having ESP even if she predicts all outcomes correctly.

"The good ones into the pot,
The bad ones into the crop."

We want to identify the skilled managers (*"The good ones into the pot"*) but exclude at the same time the lucky managers (*"The bad ones into the crop"*). But, unlike her, we must face the imperfect nature of statistical tests.[6] As a result, naïve testing, without taking the multiple evaluations into account, will allow lucky managers to creep in. This pitfall is rephrased in Grinold and Kahn (2000) in the following words:

> "The fundamental goal of performance analysis is to separate skill from luck. But, how do you tell them apart? In a population of 1,000 investment managers, about 5 percent, or 50, should have exceptional performance by chance alone. None of the successful managers will admit to being lucky; all of the unsuccessful managers will cite bad luck."

1.2 The Solution

We now discuss the solution to the challenge. In doing so, we first need to introduce some notation.

There are N funds in the universe and the (common) return history comprises T observations. The alpha of a given fund manager with respect to his corresponding benchmark is denoted by α_n, for $n = 1, \ldots, N$. The choice of the appropriate benchmark is up to the FoF manager, not the statistician. For example, the benchmark could simply be the riskfree rate. Or it could be a hedge fund index, comprised of funds that have a similar investment style. More generally, multi-factor benchmarks as in Kosowski et al. (2007) are also possible.

We look at individual hypotheses of the form:

$$H_n : \alpha_n \leq 0 \quad \text{vs.} \quad H_n' : \alpha_n > 0 \ . \tag{1.1}$$

[6] Cinderella enjoyed the help of pigeons who could perfectly tell whether a particular lentil was 'good' or 'bad'.

So for each fund, the null hypothesis corresponds to a non-skilled manager (that is, his alpha is negative or zero), while the alternative corresponds to a skilled manager (that is, his alpha is positive). The two sets of non-skilled (or potentially lucky) managers and skilled managers are denoted by $\mathcal{I}$ and $\mathcal{I}'$, respectively:

$$\mathcal{I} = \{n : \alpha_n \leq 0\} \quad \text{and} \quad \mathcal{I}' = \{n : \alpha_n > 0\} \ .$$

The goal is to make individual decisions about each testing problem (1.1) while controlling the probability of lucky managers to pass the test by chance. A particular manager n is declared skilled by our statistical method if H_n is rejected in favor of H_n'. Depending on the (unknown) state of nature, there are two possibilities if this happens. On the one hand, if H_n is actually true, we make a mistake in the sense of declaring a non-skilled manager as skilled. Or, in the lingo of the statistician, we make a *false discovery*. On the other hand, if H_n is actually false, we correctly identified a skilled manager as skilled. Or, in the lingo of the statistician, we made a *true discovery*.

1.2.1 Formal Description of the Solution

For the purpose of this paper, accounting for multiple testing means that we are concerned about the possibility of even one lucky manager to pass the test or, in other words, to make even a single false discovery.[7]

Let F denote the number of false discoveries that our statistical method is going to make. Then the *familywise error rate* (FWE) is defined as the probability of making even one false discovery:

$$\text{FWE} \equiv P\{F > 0\} = P\{\text{Reject at least one } H_n \text{ with } n \in \mathcal{I}\} \ .$$

An appropriate statistical multiple testing method then ensures that this probability lies below some small, prespecified level, say 5% or 10%. Usually this level is denoted by α in the statistical literature but here we shall denote it by δ instead in order to avoid any confusion with the α's of the fund managers. Therefore, the goal is to ensure that:

$$\text{FWE} \leq \delta \ .$$

By limiting the probability that even one lucky manager passes the test, we can in turn be confident that all managers identified by the statistical method are truly skilled. More specifically, assume $\delta = 10\%$. Then, after applying the method, we can be $1 - \delta$, or 90%, confident that all identified managers are truly skilled. As a result, with a high probability, our statistical FoF portfolio will only consist of skilled managers.

[7] Put in the context of Cinderella, we do not want even one bad one ending up in the pot.

1.2.2 Implementation of the Solution

Implementing the solution in practice is anything but trivial. A host of statistical problems arise, among others:

- The non-normality of hedge fund returns.
- The time series nature of hedge fund returns.
- The choice of the individual performance measures: raw alpha estimate $\hat{\alpha}$ vs. t-statistic. The t-statistic is obtained by dividing the raw alpha estimate by its estimation uncertainty, which is quantified via a standard error.
- Accounting for the dependency across managers in order to improve the power of the statistical method, that is, its ability to detect skilled managers.

For each fund we compute an estimate of α_n, denoted by $\hat{\alpha}_n$, and a corresponding standard error $\hat{\sigma}_n$.[8] The 'studentized' test statistic for testing H_n vs. H_n' is then given by

$$t_n = \frac{\hat{\alpha}_n}{\hat{\sigma}_n} \ .$$

The funds are ranked according to their test statistics, that is, the fund with the largest t_n statistic is the top fund according to this ranking and so on.

Alternatively, it would be possible to rank the fund managers simply according to their non-studentized test statistics $\hat{\alpha}_n$, that is, according to the 'raw' alpha estimates. While this is actually the more common approach in the mainstream finance media, we consider it misguided. Ranking by the $\hat{\alpha}_n$ does not account for the (wildly) varying risks taken on by the various fund managers. On the other hand, ranking by the t_n does, since a larger risk will be reflected by a larger standard error $\hat{\sigma}_n$. This in the very same spirit as using the Sharpe ratio (that is, a risk-adjusted performance measure) to judge the performance of a fund manager rather than the raw excess return (that is, a not-risk-adjusted performance measure).

How to compute $\hat{\alpha}_n$ and the corresponding standard error $\hat{\sigma}_n$ depends on the given benchmark. A very general setup covering most practical applications are multi-factor benchmarks as in Kosowski et al. (2007). In such cases, $\hat{\alpha}_n$ can be computed from a standard OLS time series regression, based on the observed fund return and factor data. But care must be taken in computing the standard error $\hat{\sigma}_n$. It would be generally wrong to simply use the standard error provided by the OLS output, since it does not properly account for the time series nature of hedge fund returns (and potentially also some of the factors). Instead one should use a HAC standard error[9] employing kernel estimation techniques; for example, see Andrews (1991) and Andrews and Monahan (1992).

[8] The standard error $\hat{\sigma}_n$ is an estimate of the unknown standard deviation of $\hat{\alpha}_n$.

[9] HAC stands for 'heteroskedasticity and autocorrelation consistent'.

Once the test statistics t_n have been obtained, it is the task of the multiple testing method to compute a cutoff value, denoted by d, from the joint track records of all managers in the investment universe and then declare those managers as skilled for which $t_n > d$. Crucially, this has to be done in a way such that the FWE is controlled. Of course, controlling a multiple testing criterion is only one side of the coin. It could be trivially achieved by never declaring any fund manager as skilled (that is, by choosing $c = \infty$). Naturally, there is also the other side of the coin. At same time, we wish to identify as many skilled managers as possible. So in the lingo of the statistician, we want to employ a multiple testing method with as much *power* as possible. The current state of the art is developed Romano and Wolf (2005) and can be summarized as follows.

It turns out that the ideal critical value d would be given by the $1-\delta$ quantile of the following random variable:

$$\max_{1\leq n\leq N} \frac{(\hat{\alpha}_n - \alpha_n)}{\hat{\sigma}_n} \ . \tag{1.2}$$

Importantly, the value of d is not only determined by the N marginal distributions of the individual statistics $(\hat{\alpha}_n - \alpha_n)/\hat{\sigma}_n$ but also by their cross-dependence structure. Such a procedure is not realistic, nevertheless, since the distribution of the random variable (1.2) is not known in practice. However, a consistent estimator of d, denoted by $\hat{d}$, can be obtained by a bootstrap method. Namely, $\hat{d}$ is obtained as the $1-\delta$ quantile of the following random variable:

$$\max_{1\leq n\leq N} \frac{(\hat{\alpha}^*_n - \hat{\alpha}_n)}{\hat{\sigma}^*_n} \ . \tag{1.3}$$

To this end, artificial return data are generated by an appropriate time series bootstrap mechanism. The estimator of α_n and its corresponding standard error computed from this artificial data set are denoted by $\hat{\alpha}^*_n$ and $\hat{\sigma}^*_n$, respectively. The algorithm to compute $\hat{\sigma}^*_n$ generally depends on the particular bootstrap mechanism chosen. We refer the interested reader to Romano and Wolf (2005) for the details. The *bona fide* decision rule is then to declare all funds managers as skilled for which $t_n > \hat{d}$.

The price one has to pay for replacing d by $\hat{d}$ is that control of the FWE is replaced by *asymptotic* control of the FWE:

$$\limsup_{T\to\infty} \mathrm{FWE} \leq \delta \ .$$

However, simulation studies show that for practically relevant sample sizes T, the finite-sample control of the FWE is very satisfactory; see Romano and Wolf (2005) and Romano et al. (2008).

Remark 1.2.1. A key innovation of Romano and Wolf (2005) is to develop a *stepwise* method to detect as many skilled managers as possible. Instead of using a formal algorithm, it can be quite easily described in English. Assume there are $N = 100$ managers under test simultaneously and that 10 of them are detected as skilled using the procedure described above. We are left then with a smaller universe of 90 managers. The 'trick' now is to use the same formal procedure on the remaining smaller universe, which might lead to the detection of some further skilled managers.

The reason is as follows. The individual test statistics t_n will stay the same, of course. However, the critical value $\hat{d}$ in this second step will generally be smaller, since now we are looking at the maximum over 90 statistics, rather than over 100 statistics, and so the resulting $1-\delta$ quantile will be at most as large but typically strictly smaller. So some further rejections may result. In which case we continue to play the same game in the third step and so on, until no further rejections result any more.

This more powerful stepwise method still provides asymptotic control of the FWE. ■

For the empirical analysis of this paper, we use the riskfree rate as the common benchmark for all hedge funds. In this case, the corresponding alpha is simply the expected excess return of the fund (over the riskfree rate). For a given fund, $\hat{\alpha}_n$ is computed as the sample average excess return over the observed investment period. The corresponding standard error $\hat{\sigma}_n$ is a standard HAC standard error employing a kernel estimation technique. In particular, we use the method of Andrews and Monahan (1992), based on the QS (quadratic spectral) kernel.

1.2.3 Comparison to Related Approaches

Needless to say, we are not the first ones to suggest to carry out hedge fund selection based on the managers' track records. We lack the time and the space to discuss all previously suggested approaches in detail and so limit ourselves to two selected comparisons.

Our method will, with a high probability, only identify skilled managers. As described above, the method works in the following way. Rank the fund managers by a certain performance criterion computed from their respective track records. Then based on the chosen input parameter δ, the method selects an *a priori* random number of the top funds, which are then declared as skilled. In other words, the threshold a manager must pass is actually computed from the joint track records themselves and is therefore stochastic. Knowing the number of funds in the investment universe will not tell us how many funds will end up in the FoF portfolio until we actually jointly examine all the track records.

This is in contrast to some previous approaches that suggest to pick either an *a priori* fixed percentage or or an *a priori* fixed number of the top funds for the FoF portfolio; see Joehri and Leippold (2006) and Gregoriou et al. (2006), respectively. In discussing such approaches, we will focus on the fixed-percentage strategies; the critique would be similar for the fixed-number strategies.

The obvious question is how to pick the percentage *ex ante*? When backtesting the strategy, for a given investment universe and a given investment period, there usually will be a certain percentage leading *ex post* to a very good performance. But there is no universally 'optimal' percentage. The results will vary with the investment universe and/or the investment period. To put it in the context of non-skilled vs. skilled managers and selecting two (overly) extreme scenarios just to make the point: if all managers are non-skilled, the optimal percentage is zero;

if all the managers are skilled, the optimal percentage is 100. Knowing from previous published studies that a certain percentage worked well for a certain investment universe during a certain investment period, is not overly helpful to a FoF manager faced with a different universe and a different period. In fact, such information might actually be quite misleading.

On the other hand, the use of our multiple testing methods gives the FoF manager the confidence that for his specific investment universe and investment period, the selected fund managers are all skilled. And such a selection should result in continued attractive future performance for the corresponding FoF portfolio. Whether this indeed is the case will be examined in the next section by means of some backtesting exercises. Importantly, these exercises do not require any hindsight knowledge but instead yield true 'out-of-sample' performances.

1.3 Investment Universes and Portfolio Construction

We use the CISDM database from `http://wrds.wharton.upenn.edu` and a customized Eurekahedge datafeed from `http://www.eurekahedge.com` to get monthly series of net-of-fees hedge fund returns.

We apply an 'observe ten years–invest one year' strategy with a three-month sell lag, moving at an annual frequency. More specifically, on October 1, of every year y, we feed 117 months of past return data into the multiple testing method. It then detects the statistically significantly skilled fund managers. We then invest in the equal-weighted portfolio of the detected hedge funds from January to December in year $y+1$. Then the procedure repeats, that is, on October 1 of year $y+1$, we already need to decide which hedge funds we want to invest in over the next year $y+2$. Given the annually moving 'observe ten years–invest one year' strategy, six investment periods from year 2000 to 2005 (for CISDM) and from year 2002 to 2007 (for Eurekahedge), respectively, are obtained.

At any given investment point in time, we are only selecting from a certain sub-universe of all funds contained in the respective database (CISDM or Eurekahedge). First, we restrict attention to funds which both have a complete 117-month return history *and* are open to investment at this point. Second, we exclude funds that (overall) lost money over this 117-month period.[10] Third, we exclude all funds that have at least one recorded monthly return exceeding 50% in absolute value.[11] Fourth, to avoid the inclusion of funds which are 'too similar' to each other, we impose that all the pairwise sample correlations over the 117-month period lie below 0.95, so some further funds might have to be excluded.[12]

[10] Since we are benchmarking against the risk-free rate always, no fund manager that lost money overall could possibly be considered outperforming.

[11] The motivation here is two-fold. On the one hand, such recorded returns might simply correspond to data-entry mistakes. On the other hand, even if such returns are true, they may have a large impact on the data analysis because of their undue effect on sample means, sample standard deviations, and sample Sharpe ratios.

[12] The motivation here is that sometimes 'basically the same fund' can appear under slightly different names. We implicitly take the stance that the FoF manager would only want to invest in one of such funds.

In addition to the equal-weighted portfolio of the outperforming funds, we build a global minimum variance portfolio (GMV) with the outperforming funds. Specifically, given K outperforming funds over 117 months detected by our multiple testing method, we solve the following optimization problem within each 117 months window

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}'\hat{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w} \geq \mathbf{0} \\ & \mathbf{w}'\mathbf{1} = \mathbf{1}, \end{aligned} \tag{1.4}$$

using quadratic programming methods. Since the true covariance matrix Σ is unknown, we estimate it using a suitable shrinkage estimator from the joint track records of the K outperforming funds over the last 117 months; see the Appendix for details. Optimization problem (1.4) returns an optimal weight $\mathbf{w}^*$ for each 117 month window. In the following year, one then invests in the $\mathbf{w}^*$-weighted portfolio of the outperforming funds. The equal-weighted portfolio is simply the $\mathbf{w}^* = [1/K \dots 1/K]$ weighted portfolio of the outperforming funds. The rebalancing and the three months sell-lag is as before.

As pointed out before, selecting an appropriate benchmark for a given hedge fund is the task of the FoF manager, not of the statistician. Since we are 'ignorant' in this respect, we simply chose the riskfree rate as the universal benchmark. Such a choice certainly appears reasonable and may even be the natural one from certain view points. In practice, the particular riskfree rate we use is from the CRSP Risk Free Rates file.[13]

The multiple testing criterion we employ is the control of the FWE with parameter $\delta = 10\%$. So at any given point in time, we can be 90% confident that all identified managers are truly skilled.

It is then natural to ask whether there is any 'value' in our statistical technique of constructing a FoF. An obvious competitor is the $1/N$ portfolio, that is, the equal-weighted portfolio of all available hedge funds. Recent work by DeMiguel et al. (2009), in the context of building equity portfolios, shows that this simple minded portfolio is actually surprisingly difficult to outperform for statistical methods that construct portfolios based on the past return data. However, in contrast to equity investing, the $1/N$ portfolio is often not feasible for a FoF manager, given the various minimum investments of the individual funds. Hence, it is of interest to see whether statistical FoF portfolio, based on a much smaller investment universe, can do (at least) as well as the $1/N$ portfolio. So for each investment universe, we also include the $1/N$ portfolio in our study.

Remark 1.3.1. Having a smaller investment universe by applying a multiple testing method rather than investing in all available funds is particularly important when portfolio optimization, such as choosing the global minimum variance portfolio, is used. In this case, the smallest weight (or investment portion) will often be much smaller than the inverse of the number of

[13] We employ the average rate of *ask* and *bid*.

funds to invest in. So the larger the number of funds, given the various minimum investments, the less feasible such a 'optimized' strategy becomes. ■

Furthermore, we consider two investable hedge fund indices for comparison. The HFRX Global Investable Hedge Fund Index is from `www.hedgefundresearch.com` and the CS/Tremont Investable Hedge Fund Index from `www.hedgeindex.com`. Note that the inclusion of these indices somewhat amounts to comparing apples to oranges, since they correspond to investment universes different from both the CISDM and the Eurekahedge databases. Nevertheless it is interesting to see how our statistical FoF portfolios fare against some 'real life' competitors.

1.3.1 Idealistic Setup

In a first analysis, all hedge funds that have a complete return history of 192 months are part of our chosen investment universes. This is idealistic, since we will never know in January 2000, say, which funds will survive until December 2005 in order to restrict our attention to them. Nevertheless, it is also of interest to compare our statistical FoF portfolio to the $1/N$ portfolio in this context.

Remark 1.3.2. Constructing investment portfolios based on statistical multiple testing methods, investing in assets which are established as outperforming, is certainly not restricted to the hedge fund industry. More generally, this approach could also be applied to equities, bonds, foreign exchange, etc. The frequency of individual assets 'dying' in such alternative markets will often be much reduced compared to the hedge fund industry, or even (close to) zero. So including the results for a world without dying individual funds/assets is not only of academic interest. ■

In a second step, we will make the investment setup more realistic with respect to the characteristics of the hedge fund industry and not using any future knowledge about fund survivorship.

Either way, we always impose a realistic sell lag of three months. That is, we have to decide at October 1 in year $y-1$ which funds to sell at January 1 of year y. For simplicity, we synchronize the buy decisions with the sell decisions. So on October 1 of year $y-1$, the portfolio to be held throughout year y is chosen.

Our CISDM investment universe comprises 97 hedge funds, ranging from January 1990 to December 2005. The Eurekahedge investment universe contains 61 hedge funds over the period January 1992 to December 2007. Restricting attention to the hedge funds actually open to investment throughout the 16-year period further reduces the sizes of the two universes to 91 and 54, respectively.

1.3.2 Realistic Setup

In the second part of our analysis, we evaluate a more realistic strategy, both for the FoF and the $1/N$ portfolios as follows. In October of a given year, we take as the investment universe all funds that have a complete 117-month history. As before, we impose a reasonable sell lag of three months and synchronize the sell decisions with the buy decisions.

We then construct both our statistical FoF portfolios and the $1/N$ portfolio and hold them for a year. During that year, some funds might 'die' of course. Not all funds will generally return all money to the investors. We, therefore, assume a uniform recovery rate of 90% of the investments at the time a fund closes down.[14] The recovered money is then invested in the riskfree rate for the remainder of year. Then we play the same game again next October. So in this way, the size of the investment universe N actually varies over time. Finally, we impose a disinvest-reinvest restriction, as many fund managers are not willing to tolerate a come-and-go-as-you-please behavior of investors. If we disinvest from fund n in October of any year, we are not allowed to reinvest in fund n in any of the following years anymore.[15]

The sizes of the CISDM investment universes only containing open funds are 86, 116, 160, 211, 268, 371 for the years 2000, 2001, ..., 2005, respectively. The sizes of the Eurekahedge investment universes with only open funds are 92, 118, 137, 138, 136, 119 for the years 2002, 2003, ..., 2007, respectively.

1.3.3 Statistical Significance of Portfolio Outperformance

Of course, we must keep in mind that any performance measures computed from a finite investment period are only sample-based estimates rather than true 'population numbers' (or *parameters* in the lingo of the statistician). So when comparing two portfolios based on a given performance measure, we cannot necessarily conclude that the portfolio with the higher sample-based estimate is indeed better. In other words, we cannot claim any statistical significance based on the sample-based estimates only. To this end, rather, we need to employ a proper statistical test.

Let us focus on the Sharpe ratio which, arguably, is the single most important performance measure. We want to establish whether the true 'underlying' Sharpe ratio of the statistical FoF portfolio is indeed larger than the one of the $1/N$ portfolio in the realistic setup. Denote these two parameters by $\mathrm{SR}_{\mathrm{FoF}}$ and $\mathrm{SR}_{1/N}$, respectively. Further, denote their difference by Δ, that is,

$$\Delta = \mathrm{SR}_{\mathrm{FoF}} - \mathrm{SR}_{1/N} \ .$$

Since we have an *a priori* belief that $\Delta > 0$ and would like to 'verify' this belief by a statistical

[14] Of course, recovery rates vary in practice. But this additional knowledge is not available to us. So to impose a fixed 'average rate' appears the best feasible solution.

[15] The results do not change much if this disinvest-reinvest restriction is not imposed. For the sake of brevity, the results without this restriction are not reported.

test, we consider a one-sided test of the kind:

$$H : \Delta \leq 0 \quad \text{vs.} \quad H : \Delta > 0 \ .$$

For both investment universes, the sample-based estimates $\hat{\Delta}$ are indeed positive: for the CISDM universe, we obtain $\hat{\Delta} = 0.37 - 0.32 = 0.05$; for the Eurekahedge universe, we obtain $\hat{\Delta} = 0.37 - 0.27 = 0.10$, as reported in Table 1.3. But again, this does not 'prove' that the two population Δ's are also positive.

Testing for the difference between two population Sharpe ratios is a non-trivial matter. The most commonly used method in the finance literature is the test of Memmel (2003), which is a corrected version of the earlier test of Jobson and Korkie (1981). Unfortunately, this test was derived using the overly strict assumptions of return data that follow a normal distribution and are additionally independent over time. At least one of these two assumptions is generally violated in practice. For hedge fund return data, typically both assumptions are violated. As a result, the test of Memmel (2003) tends to overstate the statistical evidence that is really contained in the observed data. Therefore, since we want to demonstrate that our FoF portfolios outperform the $1/N$ portfolios with respect to the Sharpe ratio, using the test of Memmel (2003) would actually be tempting. However, it would not be correct.

Ledoit and Wolf (2008) propose a bootstrap test that instead yields reliable inference in the presence of non-normal return distributions and time series effects. In other words, it gives a fair appraisal of the statistical significance actually contained in the observed data. Note that their bootstrap test is designed for two-sided hypotheses of the kind

$$H : \Delta = 0 \quad \text{vs.} \quad H' : \Delta \neq 0,$$

but it can be easily modified to apply to the one-sided case as well.

As stated, we believe that the Sharpe ratio is the single most important performance measure. Looking at measures that are not adjusted for the risk taken out by the fund manager, such as the average (excess) return can be quite misleading. Nevertheless, we can apply a statistical test to the difference between average (excess) returns as well. Again, we propose to use a bootstrap test that yields reliable inference in the presence of non-normal return distributions and time series effects. Testing for means is easier than testing for Sharpe ratios. Therefore, the test of Ledoit and Wolf (2008) can be 'simplified' in a straightforward manner to deal with means.

1.4 Results

The results are summarized in Tables 1.1 and 1.2 for the idealistic setup and in Tables 1.3 and 1.4 for the realistic setup, respectively. Importantly, all summary statistics are on a

monthly basis, that is, they are not annualized.[16] In addition, Figures 1.1 and 1.2 provide some graphical representation of the various return distributions.

1.4.1 Idealistic Setup

First in Table 1.1, we report the number of hedge funds making up the statistical FoF portfolio in each of the six annual investment periods. For the CISDM portfolio this number varies between 3 and 9, compared to a universe size of 91. For the Eureka portfolio, this number varies between 1 and 5, compared to a universe size of 54. The size of the HFRX index varies over time, always being larger than 60. The size of the CS/Tremont index is 60.

Second, we report the mean of the monthly excess log returns over the six annual investment periods. We find that for both investment universes (and their slightly different respective investment periods), the statistical FoF portfolios yield a lower excess return than the $1/N$ portfolio. However, these differences are not statistically significant, as reported in Table 1.2.

Third, we report the mean of the 'raw' log monthly returns (that is, not in excess of the riskfree rate). Not surprisingly, the comparisons are qualitatively very similar to the ones for the excess returns.

Fourth, we report the Sharpe ratios of the monthly log excess returns. As already stated, for both investment universes, our statistical FoF portfolios have a (somewhat) smaller excess return and a (much) smaller portfolio size than the $1/N$ portfolio. Typically, one would expect smaller portfolios to have less favorable Sharpe ratios than larger ones due to diversification effects. However, the opposite is the case for both investment universes, with the differences being rather large at times. This is especially remarkable in case of the Eurekahedge universe where the size of the statistical FoF portfolios ranges from 1 to 5. Statistical significance at the 10% level is only achieved in one case, though: namely for the EW-FoF portfolio with the Eureka data.

Fifth, we report the maximum drawdown over the out-of-sample investment period of $6 \cdot 12 = 72$ months. Again, for both investment universes, the statistical FoF portfolios outperform the $1/N$ portfolio, adding further evidence to the claim that multiple testing technique successfully identifies a small number of skilled managers from the large investment pool.

The boxplots in Figure 1.1 clearly show that the $1/N$ portfolio, despite its larger universe size, yields returns that are much more variable compared to the two statistical portfolios. In addition, portfolio optimization appears successful in the sense that the returns of GMV-FoF are somewhat less variable compared to EW-FoF.

We finally note that the statistical portfolios generally compare favorably to the investable

[16] While annualizing (excess) returns is straightforward, annualizing Sharpe ratios is not. The usual method of multiplying the monthly Sharpe rations by $\sqrt{12}$ is misleading for hedge funds due to the autocorrelation of the returns over time; see Lo (2002).

indices as well.

Table 1.1: Performance of Portfolios: Idealistic Setup

	# of hedge funds in each of the 6 years	average exc. return	average return	Sharpe ratio	maximum drawdown
CISDM data, investment period: Jan 2000 – Dec 2005.					
EW-FoF	9, 9, 3, 7, 5, 8	0.38%	0.60%	0.28	−4.22%
GMV-FoF	9, 9, 3, 7, 5, 8	0.42%	0.61%	0.59	−1.47%
$1/N$	91	0.51%	0.73%	0.20	−10.02%
HFRX Global	> 60	0.39%	0.64%	0.28	−3.92%
CS/Tremont	60	0.38%	0.60%	0.48	−2.06%
Eurekahedge data, investment period: Jan 2002 – Dec 2007.					
EW-FoF	1, 1, 1, 3, 5, 5	0.40%	0.63%	0.57	−1.89%
GMV-FoF	1, 1, 1, 3, 5, 5	0.38%	0.60%	0.64	−0.56%
$1/N$	54	0.64%	0.86%	0.31	−7.52%
HFRX Global	> 60	0.27%	0.49%	0.23	−3.57%
CS/Tremont	60	0.35%	0.57%	0.40	−2.68%

Table 1.2: Statistical Significance of Outperformance: Idealistic Setup

	Alternative hypothesis	i='CISDM'	i='Eureka'
j='mean excess return'	$\mu^{\text{exc}}_{\text{EW-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.35$	$p = 0.18$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{EW-FoF}}$	$p = 0.36$	$p = 0.09$
j='mean excess return'	$\mu^{\text{exc}}_{\text{GMV-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.38$	$p = 0.17$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{GMV-FoF}}$	$p = 0.20$	$p = 0.11$

Note: If a p-value is smaller than α, then the data supports the alternative hypothesis at significance level α.

1.4.2 Realistic Setup

First in Table 1.3, we report the number of hedge funds making up the statistical FoF portfolio in each of the six annual investment periods. We observe that the sizes of the CISDM FoF portfolios vary between 10 and 14. The Eureka FoF portfolios contain between 9 and 21 funds. The size of the HFRX index varies over time, always being larger than 60. The size of the CS/Tremont index is 60.

Second, we report the mean of the monthly excess log returns over the six annual investment periods. We see again that the mean excess monthly returns are lower than their $1/N$

counterparts. However, these differences are not statistically significant; see Table 1.4.

Third, we report the mean of the 'raw' log monthly returns (that is, not in excess of the riskfree rate). Not surprisingly, the comparisons are qualitatively very similar to the ones for the excess returns.

Fourth, we report the Sharpe ratios of the monthly log excess returns. As before, the statistical portfolios yield consistently higher Sharpe rations compared to the $1/N$ portfolio, though not at a level of statistical significance.

Fifth, we report the maximum drawdown over the out-of-sample investment period of $6 \cdot 12 = 72$ months. Again, for both investment universes, the statistical FoF portfolios outperform the $1/N$ portfolio, with the differences being rather large. In fact, for both universes, the $1/N$ portfolio has the worst drawdown of all five portfolios.

The boxplots in Figure 1.2 clearly show that the $1/N$ portfolio, despite its larger universe size, yields returns that are much more variable compared to the two statistical portfolios. In addition, portfolio optimization appears successful in the sense that the returns of GMV-FoF are somewhat less variable compared to EW-FoF.

We finally note that the statistical portfolios generally compare favorably to the investable indices as well.

Table 1.3: Performance of Portfolios: Realistic Setup

	# of hedge funds in each of the 6 years	average exc. return	average return	Sharpe ratio	maximum drawdown
CISDM data, investment period: Jan 2000 – Dec 2005.					
EW-FoF	10, 14, 13, 14, 10, 11	0.36%	0.58%	0.37	−1.83%
GMV-FoF	10, 14, 13, 14, 10, 11	0.20%	0.41%	0.33	−3.66%
$1/N$	86,116,160,211,268,371	0.54%	0.76%	0.32	−5.62%
HFRX Global	> 60	0.39%	0.61%	0.28	−3.92%
CS/Tremont	60	0.38%	0.60%	0.48	−2.06%
Eurekahedge data, investment period: Jan 2002 – Dec 2007.					
EW-FoF	18, 21, 21, 21, 10, 9	0.26%	0.48%	0.37	−3.55%
GMV-FoF	18, 21, 21, 21, 10, 9	0.30%	0.53%	0.67	−0.60%
$1/N$	92,118,137,138,136,119	0.46%	0.68%	0.27	−5.73%
HFRX Global	> 60	0.27%	0.49%	0.23	−3.57%
CS/Tremont	60	0.35%	0.57%	0.40	−2.68%

Remark 1.4.1. We generally fail to find statistical significance when testing for outperfor-

Table 1.4: Statistical Significance of Outperformance: Realistic Setup

	Alternative hypothesis	i='CISDM'	i='Eureka'
j='mean excess return'	$\mu^{\text{exc}}_{\text{EW-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.27$	$p = 0.17$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{EW-FoF}}$	$p = 0.34$	$p = 0.33$
j='mean excess return'	$\mu^{\text{exc}}_{\text{GMV-FoF}} < \mu^{\text{exc}}_{1/N}$	$p = 0.11$	$p = 0.26$
j='Sharpe ratio'	$SR_{1/N} < SR_{\text{GMV-FoF}}$	$p = 0.54$	$p = 0.11$

Note: If a p-value is smaller than α, then the data supports the alternative hypothesis at significance level α.

mance. This may not be surprising, given that it is notoriously difficult to find statistical significance in small samples of noisy financial returns (our out-of-sample period only comprises 72 months). There is, nevertheless, a clear and strong pattern. For each performance criterion (average excess return, Sharpe ratio, or maximum drawdown), there is a total of eight comparison cases (two setups, two data sets, and two statistical portfolios). In all eight cases, the story is always the same: the statistical portfolio yields a lower average excess return but outperforms the $1/N$ portfolio both in terms of the Sharpe ratio and the maximum drawdown. The latter two criteria are probably more relevant, as most FoF managers promote their ability to manage the risk in their portfolios.

We can also ask the question whether portfolio optimization via the GMV portfolio yields further benefits. Here the results somewhat less pronounced. In terms of the Sharpe ratio and maximum drawdown, the GMV statistical portfolio does better than the equal-weighted statistical portfolio in all four cases; although very closely for the Sharpe ratio of the CISDM GMV FoF in the realistic setup. ■

1.5 Conclusions

We have studied whether it is possible to construct hedge fund portfolios with attractive return properties based on the past track records of all managers in the investment universe alone. Importantly, such a strategy must not rely on any hindsight knowledge, say about which fixed percentage of top managers for a given investment universe and investment period would have worked well.

Our approach consists of comparing each manager to a given benchmark (which could be common or be allowed to vary with managers) and then to determine which managers statistically outperform their benchmark. Such managers are deemed 'skilled' and we simply go on to hold an equal-weighted or a global-minimum-variance portfolio of all skilled managers as our FoF portfolio. This process is repeated, and the portfolios thus updated, every year.

Crucially, in determining which managers statistically outperform their benchmark, one must take into account that a large number of managers are examined at the same time. In other

words, one must account for the problem of multiple comparisons (of managers against benchmark). We do this by employing some state-of-the-art statistical multiple testing methods. These methods take the non-normal return distributions and time series nature of hedge fund returns into account to properly control the chance of non-skilled managers creeping into our FoF portfolio. On the other hand, these methods are also optimized with respect to detecting as many skilled managers as possible in order to build a well-diversified portfolio.

We backtested this strategy (without using any hindsight knowledge) on two hedge fund universes. When comparing the performance of the statistical FoF portfolios to their most natural competitor, namely the $1/N$ portfolio, we found that they deliver consistent improvements both in terms of the Sharpe ratio and the maximum monthly drawdown. The return properties are also attractive when compared to two investable hedge fund indices (based on different investment universes).

While traditional approaches to construct FoFs, such as due diligence, will remain vital, we believe that statistical selection techniques based on the past track records alone can be an attractive (and cost efficient) alternative method. Of course, there is no reason not to combine these two approaches. Indeed, while clearly beyond the scope of this paper, the combination of more complex traditional approaches with statistical selection techniques might well result in the best of both worlds.

References

Please find the references of this essay at the end of this Ph.D. thesis.

1.A New Shrinkage Estimator for Σ

When estimating a covariance matrix based on (limited) past track records, one should not use the sample covariance matrix. This is especially true when the estimated covariance matrix is used for purposes of portfolio optimization. The intuitive reason is that the optimizer will latch on to the large estimation error contained in the sample covariance matrix and produce very unstable portfolios that often yield poor out-of-sample performance. This important point is discussed by Ledoit and Wolf (2003, 2004) who also offer a remedy. Namely, shrink the sample covariance matrix to a highly structured estimator, called the *shrinkage target*. Such an estimator will be biased, unlike the sample covariance matrix, but in return contain very little estimation error. Combining the two estimators via shrinkage will result in an optimal bias-variance trade-off.

Ledoit and Wolf (2003, 2004) suggest shrinkage targets for a universe of stocks: the single-factor model and the single-correlation model. But the targets have a common feature: the diagonal of the matrix is the same as the diagonal of the sample covariance matrix. As a result,

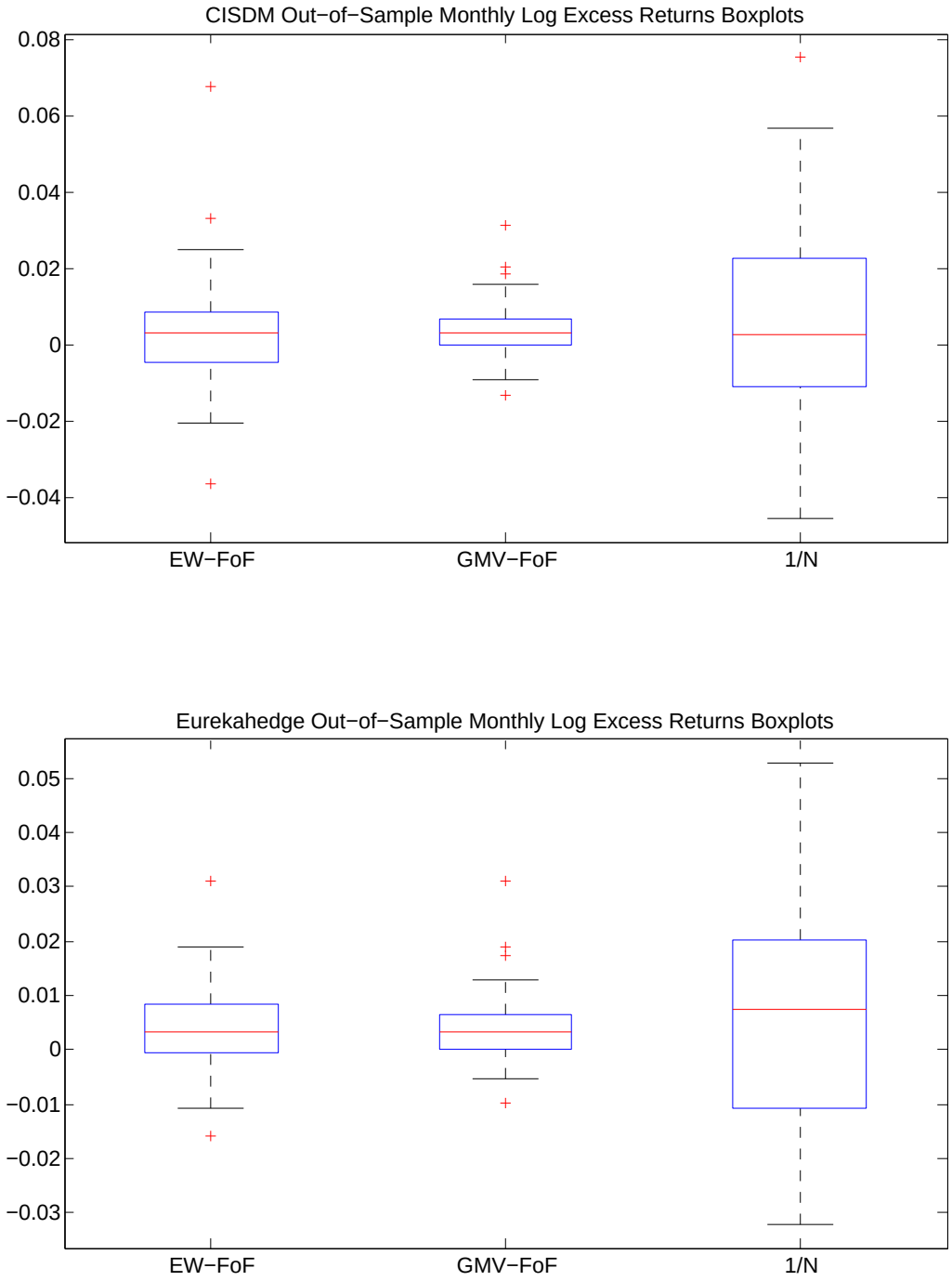


Figure 1.1: Box Plots of Out-of-Sample Log Excess Returns: Idealistic Setup

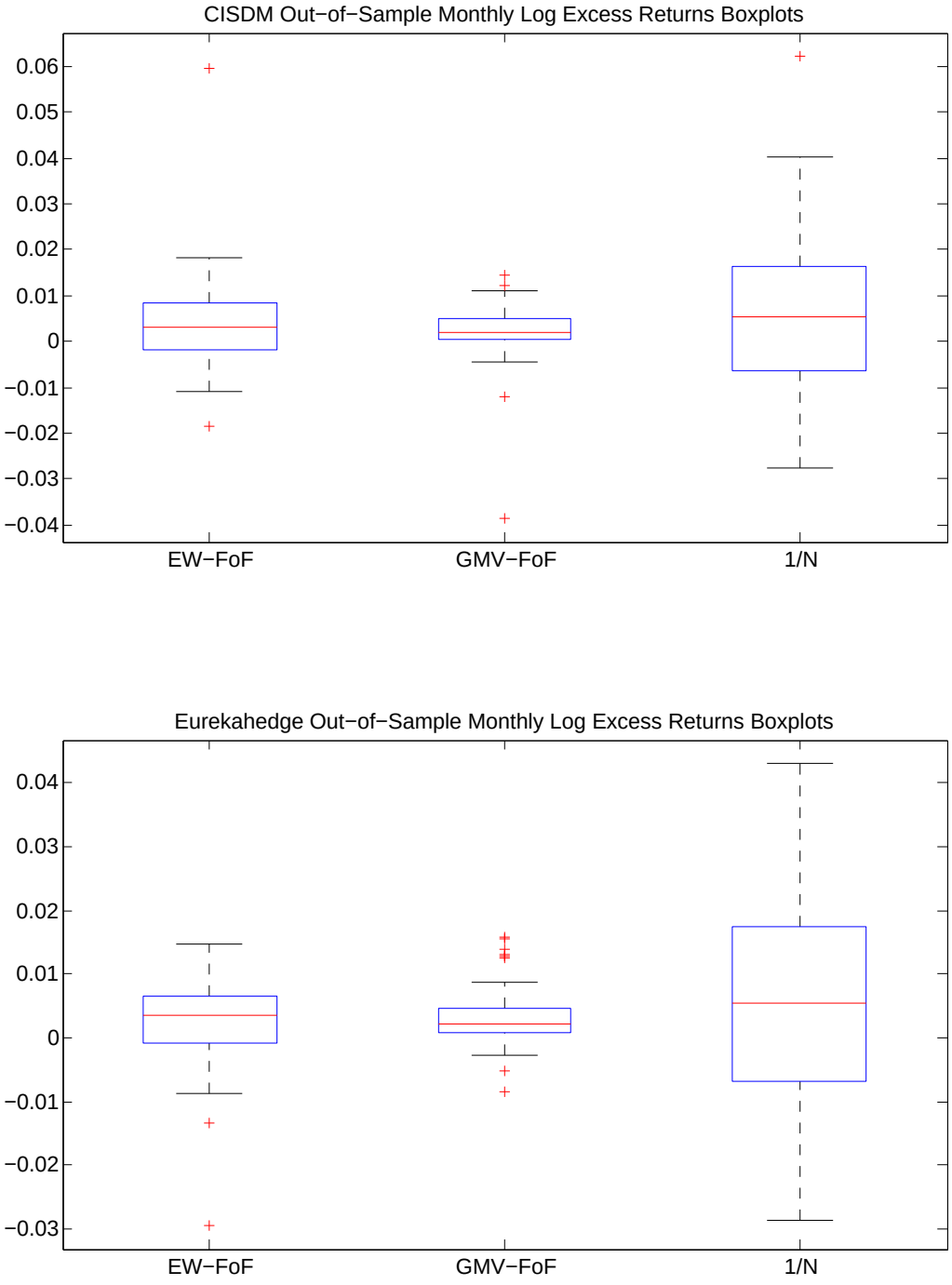


Figure 1.2: Box Plots of Out-of-Sample Log Excess Returns: Realistic Setup

only the sample covariances get shrunken/modified but not the sample variances.

We feel that such an approach is sub-optimal when dealing with hedge funds instead of stocks. Due to the wildly varying amounts of risk taken on by the various funds, already the differences between the sample variances will be overstated. It, therefore, appears useful to shrink the sample variances in addition to the sample covariances.

Therefore, we propose the *two-parameter* model as a shrinkage target. It has one common variance and one common covariance. The estimation of these two parameters is straightforward. One simply takes the average of all sample variances and the average of all sample covariances, respectively. One then is left to find a formula for the optimal shrinkage intensity. The general methodology is outlined in Ledoit and Wolf (2003) and the details are left to the reader. Computer code in the Matlab language can be downloaded for free from the following website: `http://www.econ.uzh.ch/faculty/wolf`.

Essay 2

Multiple Treatment Effects

Controlling the Danger of False Discoveries in Estimating Multiple Treatment Effects

Dan Wunderli[1]
Department of Economics, University of Zurich
Wilfriedstrasse 6, CH-8032 Zurich, Switzerland
dan.wunderli@econ.uzh.ch

December 2011

Abstract

I expose the risk of false discoveries in the context of multiple treatment effects. A false discovery is a nonexistent effect that is falsely labeled as statistically significant by its individual t-value. Labeling nonexistent effects as statistically significant has wide-ranging academic and policy-related implications, like costly false conclusions from policy evaluations. I reexamine an empirical labor market model by using state-of-the art multiple testing methods and I provide simulation evidence. By merely using individual t-values at conventional significance levels, the risk of labeling probably nonexistent treatment effects as statistically significant is unacceptably high. Individual t-values even label a number of treatment effects as significant, whereas multiple testing indicates false discoveries in these cases. Tests of a joint null hypothesis such as the well-known F-test control the risk of false discoveries only to a limited extent and do not optimally allow for rejecting individual hypotheses. Multiple testing methods control the risk of false discoveries in general while allowing for individual decisions in the sense of rejecting individual hypotheses.

KEY WORDS: False discoveries, multiple error rates, multiple treatment effects, labor market

JEL CLASSIFICATION NOS: C12, C14, C21, C31, C41, J08, J64.

[1]Many thanks to Rafael Lalive, Ashok Kaul, Rainer Winkelmann, and Michael Wolf for their comments. Thanks also go to Gregori Baetschmann, Alexandru Popescu, and participants of the microeconometrics research seminar at the University of Zurich for useful comments.

2.1 Introduction

I put the danger of false discoveries into perspective by providing simulation evidence and by reexamining treatment effects within an empirical labor market model of Lalive et al. (2005). A false discovery is a nonexistent effect that is falsely labeled as statistically significant by its individual t-value. I provide evidence that the risk of making false discoveries is unacceptably high if one does not account for the danger of false discoveries. It is shown that individual t-values even label a number of treatment effects as statistically significant that are probably false discoveries. As usual in inferential statistics, one can only 'prove beyond a reasonable doubt' that an effect exists. One can show by multiple testing methods that some individually significant treatment effects are probably nonexistent, as quantified in a so-called multiple significance level. In this paper, 'nonexistent treatment effects' should be understood in this inferential way.

In the empirical labor market model of Lalive et al. (2005), there are multiple treatment effects. These treatment effects are potentially interrelated, thus one must consider the treatment effects jointly. The central empirical question is whether some of the treatment effects being individually significant are false discoveries in the sense of not having controlled the risk of labeling nonexistent treatment effects as statistically significant, that is, of making false discoveries. I reexamine the empirical model of Lalive et al. (2005) with respect to making false discoveries. I control the risk of labeling nonexistent treatment effects as (individually) significant by using the powerful multiple testing methods from Romano and Wolf (2005).

To ease understanding, I chose the somewhat vague phrase 'one must consider the effects **jointly**' instead of the technically correct 'one must consider effects with a **multiple** error type one'. In reading jointly, most readers probably think of an F-test. However, an F-test cannot control the risk of labeling some nonexistent treatment effects as statistically significant. To see why, let us first consider the difference between testing a number of individual hypotheses and testing one joint null hypothesis; the difference between the former and multiple testing is explained in a second step. The crucial first point is that a joint null hypothesis does not allow for individual decisions in the sense of rejecting individual null hypotheses from a joint point of view.

Number of individual hypotheses versus one joint hypothesis For the sake of exposition, consider the regression

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i. \tag{2.1}$$

Suppose that $\beta_1, \beta_2, \beta_3$ measure treatment effects. It is clear that testing one **joint** null hypothesis $H_{0,joint} : \beta_1 = \beta_2 = \beta_3 = 0$ with an F-test does not necessarily yield the same result than testing **three single** null hypotheses $H_{0,1} : \beta_1 = 0$, $H_{0,2} : \beta_2 = 0$, and $H_{0,3} : \beta_3 = 0$ in the following sense. It may be the case that the first and third individual hypothesis $H_{0,1}$, $H_{0,3}$ are rejected, although the joint $H_{0,joint}$ cannot be rejected. Nonetheless, some coefficients

may still be significant by jointly considering the three coefficients, despite the fact that the F-test could not reject $H_{0,joint} : \beta_1 = \beta_2 = \beta_3 = 0$. Furthermore, it may happen that no individual t-test rejects while the joint F-test rejects, as for highly correlated regressors. By t-testing I mean calculating the $t$istic and comparing it to the appropriate quantile of the t (or normal) distribution. Of course, simply relying on individual t-tests of single null hypotheses and bluntly discarding the joint point of view of the F-test is not the solution. The more general Wald, Lagrange Multiplier (LM), or Likelihood Ratio (LR) tests have the same drawback than the F-test. They can just test one **joint** nonlinear hypothesis $H_{0,joint} : \mathbf{a}(\theta_0) = \mathbf{0}$ against its joint alternative hypothesis[2]. We really want a test that considers statistical significance jointly, as in the F-test. But we also want this joint test to tell us which individual coefficients out of the joint null hypothesis are individually significant while taking into account their joint nature. Section 2.3 provides graphical illustrations in this respect. We want a joint test in which individual rejections are possible.

Individual testing versus multiple testing To this end, multiple testing methods tell us exactly **which** null hypotheses out of the family of individual null hypotheses can be rejected at a given **multiple** significance level. A **multiple** significance level takes account of the danger of false discoveries. None of the aforementioned tests of the joint null hypothesis can optimally tell us **which single** coefficients are statistically significant as seen from a point of view joint with the other coefficients under scrutiny. Table 2.1 explains individual versus joint versus multiple testing for the case of $p > 1$ regression coefficients.

	Risk of false discoveries controlled	Individual decision possible	Number of null hypotheses
Individual tests	No	Yes	$p > 1$[a]
F-test or Wald, LM, LR	Maybe	No[b]	one[c]
Multiple testing	Yes	Yes	$p > 1$[d]

[a]There is one null hypothesis for each of the p coefficients $H_{0,s} : \beta_s = 0$ for $s = 1, \ldots, p$.
[b]Rectangular approximations to the elliptic joint confidence region are possible. See Section 2.3.
[c]All individual coefficients are merged to one **joint** null hypothesis $H_{0,joint} : \beta_1 = \ldots = \beta_p = 0$.
[d]All individual coefficients are merged to a **family** of null hypotheses $\{H_{0,s} : \beta_s = 0,\ s = 1, \ldots, p\}$.

Table 2.1: Multiple Testing allows joint testing while individual decisions are possible

The crucial point is that only multiple testing methods can generally control the risk of false discoveries, while optimally allowing for rejecting individual hypotheses from a joint point of view.

In our stylized regression example (2.1), suppose as before that only β_1 and β_3 are individually significant according to t-testing β_1, β_2, and β_3 at the 5% level. Regardless of the outcome of testing the joint $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at the 5% level, it may well be the case that a multiple

[2]Where $\mathbf{a} : \mathbb{R}^p \to \mathbb{R}^p$ is a continuously differentiable function

error type one is very large instead of being at a conventional level between 1% and 10%. In this paper, I consider a multiple error type one called the familywise error rate FWE that is defined as follows

$$\begin{aligned} \text{Familywise error rate FWE} &= P(\text{Number of falsely rejected null hypotheses} \geq 1) \\ &= P(\text{Number of false discoveries} \geq 1). \end{aligned} \tag{2.2}$$

where $P(\cdot)$ denotes the probabilty mechanism. The well-known error type one 'α' of an individual test is the probability of falsely rejecting the null hypothesis $H_0 : \beta_s = 0$ for one specific $s \in \{1, \ldots, p\}$. Thus, it makes sense that the multiple error type one FWE is an error probability as well. Controlling the error probability FWE at the 5% level means ensuring that FWE $\leq 5\%$. If the familywise error rate is not explicitly taken care of, it may easily be the case that only FWE $\leq 50\%$ holds. Section 2.2 provides a Monte Carlo simulation in this respect.

Note that a familywise error rate FWE at 50% instead of at 5% renders statistical inference useless. One cannot trust in the comparison of t-values to the conventional critical value $c_{1-\alpha/2}^{N(0,1)}$, or to its bootstrap version $c_{1-\alpha/2}^{boot}$, in presence of such a high probability of labeling one or more nonexistent effects as statistically significant. In the empirical results of this paper, we will even see an empirical instance where carrying out six individual t-tests at the 5% level translates to the FWE being around 90%, that is, the probability of having made one or more false discoveries is around 90%.

Why False Discoveries Matter A high familywise error rate FWE translates into a high risk of having labeled some treatment effects as individually significant that do not exist, which statisticians call false discoveries. The larger the family of individual null hypotheses is that are scrutinized jointly, the higher is the risk of labeling some treatment effects as significant that do not exist. Thus, one runs the uncontrolled danger of so-called false discoveries by testing merely individually.

Not controlling for false discoveries has wide-ranging academic and policy-related implications. In policy evaluation, if effects are falsely labeled as significant, wrong policies may be pursued, leading to a waste of public funds or to an unexpected deterioration where an improvement was expected. False discoveries are sometimes even published results as if there had been no prescreening of results based on individual p-values. Heckman et al. (2010) refer to this problem as 'cherry picking'. In this sense, results that 'did not work' should be reported along with results that 'worked'. The common robustness checks that report results of different specifications certainly are steps in the right direction.

The remainder of this paper is organized as follows. Section 2.2 provides simulation evidence on how large the danger of false discoveries can be by testing merely individually. Section 2.3 explains the difference between testing one joint null hypothesis and multiple testing of a family of hypotheses with some graphs. Section 2.4 briefly summarizes the labor market model in Lalive et al. (2005) that I reexamine. A more detailed description of the model is in Appendix

2.A. Section 3.4.3 describes different extents of detail in distinguishing treatment effects and reports results from individual and multiple testing of these treatment effects. Section 3.4.3 thus indicates to which extent the danger of false discoveries is ignored by individual t-tests within the empirical labor market model of Lalive et al. (2005). Section 2.6 concludes.

2.2 Individually Testing $p > 1$ Null Hypotheses versus Multiple Testing

2.2.1 Simulation Setup

The point that I illustrate in this section is: How large can the probability of falsely rejecting one or more null hypotheses (FWE) be by naively testing all p null hypotheses merely individually?

Consider the following simulation setup. There are p explanatory random variables $X_1, \ldots, X_p$, with which one associates p treatment effects $\beta_1, \ldots, \beta_p$ onto the random response variable Y. The explanatory variables may be correlated with each other. The data generating process is

$$Y = c + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon_t, \tag{2.3}$$

where $\epsilon_t \sim N(0, 1)$. There is not a single (nonzero) treatment effect: $\beta_1 = \ldots = \beta_p = 0$.

The empiricist only observes data sets of size N from the data generating process (2.3), resulting in estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ that may assume nonzero values based on an observed sample. Given that there does not exist any treatment effect, a test that labels any $\hat{\beta}_s$, $s = 1, \ldots, p$, as statistically significant commits an error type one; any sensible test at level α ensures that this error does not occur more often than $\alpha \cdot M$ times, at least asymptotically as the number of data sets $M \to \infty$. Let us check if the conventional t-test at level α falsely labels nonexistent treatment effects as significant in no more than $\alpha \cdot M$ cases, for the null hypothesis $H_{0,s} : \beta_s = 0$ versus alternative $H_{A,s} : \beta_s \neq 0$ for each $s = 1, \ldots, p$.

First, I generate $M = 2000$ data sets of size N from the data generating process (2.3), denoted as $x^{(1)}, \ldots, x^{(M)}$. Second, I compute the estimates $\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_p^{(m)}$ based on each generated data set $x^{(m)}$, as well as the respective t-statistics $t_1^{(m)}, \ldots, t_p^{(m)}$. $H_{0,s}$ is rejected on data set $x^{(m)}$ if $|t_s^{(m)}| > c_{1-\alpha/2}$ holds. If one or more t-tests reject on data set $x^{(m)}$, which is the case if the number of rejections $\sum_{s=1}^{p} \mathbf{1}[|t_s^{(m)}| > c_{1-\alpha/2}]$ is larger than one, a familywise error $FErr^{(m)}$ is committed on data set $x^{(m)}$

$$FErr^{(m)} = \mathbf{1}\left[\sum_{s=1}^{p} \mathbf{1}[|t_s^{(m)}| > c_{1-\alpha/2}] > 1\right] \tag{2.4}$$

The estimated probability of falsely labeling one or more $\hat{\beta}_s$ as statistically significant (com-

mitting a $FErr$) at level α is the arithmetic mean over all M simulation runs

$$FWE = \frac{1}{M}\sum_{m=1}^{M} FErr^{(m)} \tag{2.5}$$

Clearly, FWE $\approx \alpha$ should hold if individually t-testing p treatment effects should be any help in controlling the danger of labeling nonexistent treatment effects as statistically significant. I check this in a number of cases. Namely, let the number of (nonexistent) treatment effects p lie in $\{2, 5, 10, 20\}$. I allow the explanatory variables $[X_1, \ldots, X_p]'$ to be correlated with each other, according to the covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{p-1} \\ \rho & 1 & \rho & \ldots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \ldots & 1 \end{pmatrix}, \quad \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma}) \tag{2.6}$$

for $\rho \in \{0.9, 0.5, 0, -0.5, -0.9\}$. That is, the further the indices s, q of two explanatory variables X_s, X_q are apart, the less X_s, X_q are correlated with each other in absolute value. A data set of explanatory variables $x^{(m)}$ can then be simulated from a set of p i.i.d. $N(0, 1)$ variables by using the triangular Cholesky factor $\mathbf{C}$ as in $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}'$.

2.2.2 Simulation Results

Table 2.2 summarizes the results from simulation setup 2.2.1 for: numbers of (nonexistent) treatment effects $p \in \{2, 5, 10, 20\}$, correlation $\rho \in \{0.9, 0.5, 0, -0.5, -0.9\}$, size of data sets $N = 1000$ [3], level of individual tests $\alpha = 5\%$, and number of simulation runs $M = 2000$.

FWE	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0$	$\rho = -0.5$	$\rho = -0.9$
$p = 2$	7.3%	9.6%	10.1%	9.1%	7.1%
$p = 5$	19.4%	21.4%	23.2%	21.6%	19.8%
$p = 10$	36.4%	38.9%	40.3%	38.3%	36.7%
$p = 20$	57.8%	60.2%	63.2%	60.1%	57.9%

Table 2.2: Probability of falsely labeling one or more nonexistent treatment effects out of p as statistically significant by doing p t-tests individually at level $\alpha = 5\%$

First, note that the higher the number of treatment effects p, the higher is the probability of discovering one or more (nonexistent) treatment effects. Hence, the higher the number of multiple treatment effects is, the higher is the danger of making one or more false discoveries. Note that even for only two treatment effects $p = 2$, the FWE is 7.1% at best instead of

[3] For $N = 100$, the results are virtually identical

$\alpha = 5\%$, which a sensible (multiple) test ought to ensure. For $p = 5$ and larger, the probability of falsely labeling one or more nonexistent effects as significant rises to 63.2%.

Second, observe that the probability of labeling nonexistent treatment effects as significant (FWE) is highest for uncorrelated treatment effects $\rho = 0$. The higher the correlation between the p explanatory variables is in absolute value, the lower is the probability of committing a familywise error, which is not surprising. In case of perfect correlation $\rho = 1$, one essentially tests only one null hypothesis in individually testing all p null hypotheses $\{H_{0,s} : \beta_s = 0,\ s = 1, \ldots, p\}$; knowing one $\hat{\beta}_s$ means knowing all other $\hat{\beta}_q$, $s \neq q$. For uncorrelated $X_1, \ldots, X_p$, however, $\hat{\beta}_s$ is unrelated to $\hat{\beta}_q = 0$ for $s \neq q$: thus each single $\hat{\beta}_s$, $s = 1, \ldots, p$, poses a danger of making a false discovery.

Third, there does not seem to be a systematic pattern with respect to positive or negative correlation ρ, the latter meaning alternating between negative and positive correlation as in row $[1, \rho, \rho^2, \ldots, \rho^{p-1}]$ of $\mathbf{\Sigma}$ [4].

Bear in mind that the probability of labeling one or more nonexistent treatment effects as significant can be much higher for a given number of treatment effects p than in this simple simulation study. The empirical part provides an example where the FWE is 90% for $p = 6$ treatment effects.

This section illustrated that individual confidence intervals at level α do not control the probability of labeling one or more treatment effects at level α. Hence, the confidence intervals resulting from naively testing $s = 1, \ldots, p$ individual null hypotheses under $N(0, 1)$ or under bootstrapping

$$\text{naive } CI_{\beta_s}^{N(0,1)} = [\hat{\beta}_s - \hat{\sigma}_{\beta_s} c_{1-\alpha}^{N(0,1)},\ \hat{\beta}_s + \hat{\sigma}_{\beta_s} c_{1-\alpha}^{N(0,1)}] \tag{2.7}$$

$$\text{naive } CI_{\beta_s}^{boot} = [\hat{\beta}_s - \hat{\sigma}_{\beta_s} c_{1-\alpha}^{boot},\ \hat{\beta}_s + \hat{\sigma}_{\beta_s} c_{1-\alpha}^{boot}] \tag{2.8}$$

are too small. In geometric terms, this being too small means that the rectangular region spanned by all $s = 1, \ldots, p$ individual naive $CI_{\beta_s}^{N(0,1)}$ or naive $CI_{\beta_s}^{boot}$ has coverage probability smaller than $1 - \alpha$.

2.3 Tests of One Joint Null Hypothesis versus Multiple Testing

One of the key issues in comparing tests of one joint null hypothesis to multiple testing is whether individual decisions in the sense of rejecting individual null hypotheses are possible, as summarized in Table 2.1. From a purely mechanical point of view, individual decisions are possible with an F-test, say, by projecting the ellipsoidal confidence region onto the axes. However, the $H_{0,joint}$-derived individual confidence intervals are too large. In multiple testing, however, one sets up a rectangular joint confidence region, leading to small individual confidence intervals with favorable size and power properties.

[4] For all off-diagonal elements in $\mathbf{\Sigma}$ being negative, the Cholesky factor does not exist because $\mathbf{\Sigma}$ is not positive semi-definite in this case.

To see why, consider the case of a joint confidence region for the two-dimensional parameter (β_1, β_2) from a linear regression such as in equation (2.1), representing two treatment effects. In multiple testing, one constructs a rectangular joint confidence region, as depicted in Figure 4.1. Thus, individual decisions based on this rectangular joint confidence region are straight forward. One just projects each side of the rectangle onto the corresponding axis, where small confidence intervals result. Specifically, one finds one single multiple critical value $c_{1-\alpha}^{MTest}$ in multiple testing. The individual confidence intervals directly inferred from the rectangular joint confidence region are

$$\text{individual } CI_{\beta_s}^{MTest} = [\hat{\beta}_s - \hat{\sigma}_{\beta_s} c_{1-\alpha}^{MTest},\ \hat{\beta}_s + \hat{\sigma}_{\beta_s} c_{1-\alpha}^{MTest}] \tag{2.9}$$

for each individual parameter β_s. In the β_1, β_2 example above, the resulting $CI_{\beta_1}^{MTest}$ and $CI_{\beta_2}^{MTest}$ are large enough, such that the familywise error rate FWE as in (2.2) is controlled. But $CI_{\beta_1}^{MTest}$ and $CI_{\beta_2}^{MTest}$ are also small enough so that a rejection occurs with a high probability when $H_{0,s} : \beta_s = 0$ is wrong, meaning that the test has high statistical power.

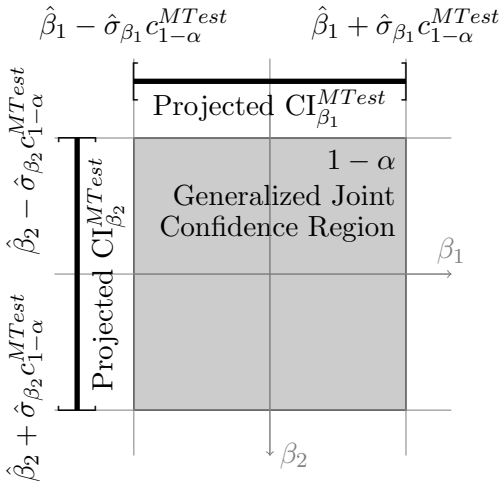


Figure 2.1: Individual CI's inferred from rectangular joint confidence region as in Romano and Wolf (2005)'s multiple testing method

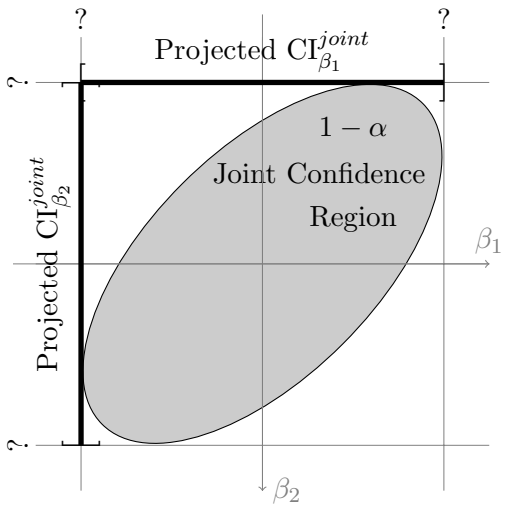


Figure 2.2: Individual CI's inferred from ellipsoidal joint confidence region as in a test of one joint null hypothesis, e.g. F-test.

In setting up a joint null hypothesis $H_{0,joint} : \beta_1 = \beta_2 = 0$ and F-testing it at significance level α, we implicitly set up an ellipsoidal joint confidence region as depicted in Figure 2.2. The more general Wald, LM, LR tests make use of ellipsoidal confidence regions as well, by considering an ellipsoidal confidence region at level $1 - \alpha$ whose ellipsoidal boundary is characterized by points with a constant $\chi^2_{1-\alpha}$ value. Testing $H_{0,joint}$ can be carried out by checking whether the null-hypothesized values of (β_1, β_2) lie outside the ellipsoidal joint confidence region, in which case one rejects $H_{0,joint}$. If we want to make individual decisions based on this $H_{0,joint}$-derived ellipsoidal confidence region, we essentially construct two individual confidence intervals from the joint confidence region. One general way of carrying out this projection is to choose the

bounds of the individual confidence intervals as the maximum value the joint confidence region attains in the associated dimension, which is represented by the light-gray lines enveloping the ellipsoid in Figure 2.2[5].

The problem resulting from these individual decisions based on $H_{0,joint}$ is the following. The ellipsoid itself has good coverage properties and is powerful with respect to $H_{0,joint} : \beta_1 = \beta_2 = 0$. However, the projected individual $H_{0,joint}$-derived confidence intervals $CI_{\beta_1}^{joint}$ and $CI_{\beta_2}^{joint}$ are too large, in the sense that they have too little power against the individual hypotheses. Additionally, it is not clear whether using $CI_{\beta_1}^{joint}$ and $CI_{\beta_2}^{joint}$ controls the familywise error rate FWE at level α.

2.4 False Discoveries in Estimating Multiple Treatment Effects

I now turn to reexamining the empirical labor market model of Lalive et al. (2005) with respect to false discoveries. For reasons of brevity, I describe their model only shortly in this section. Their competing risks duration model is outlined in more detail in Appendix 2.A.

Lalive et al. (2005) analyze the treatment effects of benefit sanctions and corresponding warnings on the duration of unemployment. Having controlled for observable characteristics x of an individual, let δ_1 denote the treatment effect of a warning, and let δ_2 denote the treatment effect of a benefit sanction on the duration of unemployment.

Common sense suggests that one tries harder to find a job when one is faced with less unemployment benefits. A warning of an unemployment benefit sanction may even be enough to induce the unemployed to look harder for jobs. If the benefit sanction is even enforced, there is a strong disutility from being unemployed, thus even the most recalcitrant unemployed have a strong incentive to find a job. This is formalized in the theoretical economic model of Boone and van Ours (2006). Lalive et al. (2005) quantify the two treatment effects of a warning, $\hat{\delta}_1$, and of a benefit sanction, $\hat{\delta}_2$, on the duration of unemployment. They use a large and reliable data set with 10,404 observations from Swiss labor market authorities to estimate the model.

In what follows, I look at cases in which the danger of making false discoveries is present. Lalive et al. (2005) do not take account of the danger of false discoveries, since they just report individual t-values. This corresponds to individually t-testing each coefficient. Given 10,404 observations, it seems reasonable to compare the observed t-statistics to quantiles from the standard normal distribution. The more coefficients that one tests individually, the higher is the danger of labeling nonexistent treatment effects as statistically significant, hence making false discoveries. Therefore, I will progressively increase the number of treatment coefficients under multiple statistical scrutiny within each of the following subchapters. In this step-wise manner, I will quantify the risk of making one or more false discoveries. Hence, we will see to

[5]The Scheffé (1953) method, based on inferring from an elliptical joint region onto individual hypotheses, is not optimal for FWE control either; see Essay 3 and Essay 4.

which extent individual t-testing ignored the risk of labeling nonexistent treatment effects as statistically significant in the empirical context of Lalive et al. (2005)'s labor market model.

2.5 Empirical Results

A multiple test of the two treatment effects $\hat{\delta}_1$ (warning) and $\hat{\delta}_2$ (benefit sanction) as in (2.10) is a natural starting point

$$\{H_{0,1} : \delta_1 = 0, \quad H_{0,2} : \delta_2 = 0\}. \tag{2.10}$$

The difference between individual and multiple testing is not pronounced though, thus the results are not reported. This may be due to the low number of two null hypotheses or due to the correlation of the test statistics in the two null hypotheses, as I illustrated in Section 2.2.

2.5.1 Duration Dependence in Treatment Effects

Family of Null Hypotheses

If one accounts for flexible duration dependence of the treatment effects themselves, individual testing may find too many individually significant treatment effects. Table A.2 of Lalive et al. (2005) reports results of treatment effects taking account of duration dependence therein. Thus, the first multiple test scrutinizes the following family of null hypotheses while controlling the risk of false discoveries

$$\begin{aligned} \{H_{0,1} : \delta_{1,\text{'0-29 days'}} = 0, \quad H_{0,2} : \delta_{1,\text{'}\geq\text{30 days'}} = 0, \\ H_{0,3} : \delta_{2,\text{'0-59 days'}} = 0, \quad H_{0,4} : \delta_{2,\text{'}\geq\text{60 days'}} = 0\} \end{aligned} \tag{2.11}$$

$\delta_{1,\text{'0-29 days'}}$ denotes the treatment effect of a warning on the duration of unemployment during the first 29 days after the warning. 30 days or more after the warning was communicated to the unemployed, $\delta_{1,\text{'}\geq\text{30 days'}}$ denotes the incremental effect to $\delta_{1,\text{'0-29 days'}}$. Hence, $\delta_{1,\text{'}\geq\text{30 days'}} < 0$ does not mean that the effect of the warning after 30 days was to increase the duration of unemployment. It means that the overall effect of a warning after 30 days is $\delta_{1,\text{'0-29 days'}} + \delta_{1,\text{'}\geq\text{30 days'}}$, which is thus smaller than the treatment effect of a warning during the first 29 days $\delta_{1,\text{'0-29 days'}}$. Therefore, the family of alternative hypotheses to (2.11) needs to be two-sided as follows

$$\begin{aligned} \{H_{A,1} : \delta_{1,\text{'0-29 days'}} \neq 0, \quad H_{A,2} : \delta_{1,\text{'}\geq\text{30 days'}} \neq 0, \\ H_{A,3} : \delta_{2,\text{'0-59 days'}} \neq 0, \quad H_{A,4} : \delta_{2,\text{'}\geq\text{60 days'}} \neq 0\}. \end{aligned} \tag{2.12}$$

Testing Results

Table 2.3 summarizes the results from individual and multiple testing of (2.11) with two-sided alternative hypotheses. Each coefficient Coeff. in the first column is followed by its estimate

$\widehat{\text{Coeff.}}$ and its t-statistic $|t|$. The columns labeled as individual tests list the critical values $c^{\cdot}$, the p-values $p^{\cdot}$, and if the individual test rejected at the 5% level, denoted as rej$^{\cdot}$. First under the normal distribution as in naive $CI^{N(0,1)}_{\beta_s}$ in (2.7), denoted as $c^{\mathcal{N}}_{0.975}$, $p^{\mathcal{N}}$, rej$^{\mathcal{N}}$. Second under bootstrapping as in naive $CI^{boot}_{\beta_s}$ in (2.8), denoted as $c^{boot}_{|\cdot|,0.95}$, p^{boot}, rej^{boot}. Details of the one- and two-sided bootstrapping methodology are in Appendix 2.B. Note that all tests are at the 5% level[6].

The columns labeled as 'multiple test' contain the multiple critical value $c^{MTest}_{|\cdot|,0.95}$ such that the familywise error rate FWE defined in (2.2) is controlled at the 5% level. This means that we control the risk of labeling one or more nonexistent treatment effects as statistically significant (making one or more false discoveries) at the 5% level. The last column rej^{MTest} indicates which individual null hypotheses can be rejected using the multiple test that controls the risk of making false discoveries.

			individual tests at 5% level						multiple test 5% level	
Coeff.	$\widehat{\text{Coeff.}}$	$\lvert t\rvert$	$c^{\mathcal{N}}_{0.975}$	$p^{\mathcal{N}}$	rej$^{\mathcal{N}}$	$c^{boot}_{\lvert\cdot\rvert,0.95}$	p^{boot}	rej^{boot}	$c^{MTest}_{\lvert\cdot\rvert,0.95}$	rej^{MTest}
$\delta_{1,\text{'0-29 days'}}$	0.4103	5.51	1.96	0.000	yes	2.00	0.000	yes	2.40	yes
$\delta_{1,\text{'}\geq\text{30 days'}}$	−0.2522	3.25	1.96	0.000	yes	2.05	0.012	yes	2.40	yes
$\delta_{2,\text{'0-59 days'}}$	0.2925	2.47	1.96	0.007	yes	2.17	0.032	yes	2.40	yes
$\delta_{2,\text{'}\geq\text{60 days'}}$	0.0437	0.42	1.96	0.338	no	1.97	0.701	no	2.40	no

Table 2.3: Results from testing four treatment effects under two-sided alternatives

First, observe that the first three coefficients are statistically significantly different from zero under the $\mathcal{N}$-distribution assumption, under bootstrapping, and under multiple testing. Thus, despite the fact that Lalive et al. (2005) just t-tested individually, they seem not having made false discoveries in the sense of having labeled nonexistent treatment effects as significant[7].

Second, note that the $\mathcal{N}$-distribution assumption seems quite accurate, given that the bootstrap critical values $c^{boot}_{|\cdot|,0.95}$ and the $\mathcal{N}$ derived $c^{\mathcal{N}}_{0.975}$ are quite close to each other. This does not hold for the associated p-values $p^{\mathcal{N}}$ and p^{boot}, though. Also note that the individual critical values $c^{boot}_{|\cdot|,0.95}$ differ from each other, while the critical value resulting from multiple testing is one and the same by construction for all coefficients.

It is possible to control more liberal multiple error types I, such as the 2-FWE being the probability of making two or more false discoveries; the interested reader finds these testing results in Appendix 2.C.

[6] $c^{\mathcal{N}}_{0.975} = c^{\mathcal{N}}_{|\cdot|,0.95}$ corresponds to the $c^{boot}_{|\cdot|,0.95}$ notation in Romano and Wolf (2005)

[7] We cannot match their reported results exactly, because we do not have the set of regressors k denoting public employment service dummies (PES), see Appendix 2.A.

Quantifying the Risk of Making False Discoveries by Individual Testing

The bootstrapping of the model produced slightly different individual critical values for the t-tests than by using the $\mathcal{N}$-assumption. That the difference between assuming $\mathcal{N}$ and bootstrapping is not pronounced is not surprising: there are $10,404$ data points. This conclusion was drawn in Table 2.3 by comparing the individual $\mathcal{N}$-assumed critical values 1.96 with the individual bootstrap critical values $c^{boot}_{|\cdot|,0.95}$ that range from 1.97 to 2.17.

How large is the risk of making one or more false discoveries[8] by individual testing? It turns out that one needs to be ready to run the risk of making one or more false discoveries at around 20% to justify the $\mathcal{N}$-assumed critical value of 1.96. Thus, by individually testing at the 5% significance level under $\mathcal{N}$, the implicit risk of making one or more false discoveries is around 20%. This is clearly too large a multiple error type one to justify it with conventional significance levels. Hence, conventional significance levels at the individual testing level do not translate into putting conventional thresholds on multiple error types one that take account of the danger of false discoveries.

Note: In principle, this high risk of making false discoveries can be attributed to the following two sources

1. Individual bootstrap testing versus multiple testing
2. Assumed normality versus data-driven approximation of the data generating distribution by bootstrapping

Given that there are $10,404$ i.i.d. data points, the sampling distribution should be close to the normal distribution as a central limit theorem suggests. Thus, the main source of this high risk of making false discoveries should be the naive individual testing instead of the multiple test. Quantifying the importance of these two sources exactly requires a very computing-intensive two-stage bootstrap simulation, as shortly described in the note at the end of Appendix 2.B.

2.5.2 Qualification Dependence in Treatment Effects

Family of Null Hypotheses

A natural question to ask is whether the effect of warnings or benefit sanctions on the duration of unemployment depends on the qualification of an unemployed person. One may also expect a systematic pattern in treatment effects with respect to gender, age groups, or with respect to other explanatory variables. I choose to differentiate the two treatment effects δ_1, δ_2 with respect to three levels of qualification $quali1$, $quali2$, $quali3$ here, resulting in six treatment

[8] That is, of falsely labeling one or more treatment effects that do not exist as statistically significant

effect coefficients. The family of null hypotheses containing the six treatment effects is

$$\begin{aligned}\{H_{0,1} : \delta_{1,\text{'quali1'}} = 0, \;\; H_{0,2} : \delta_{1,\text{'quali2'}} = 0, \;\; H_{0,3} : \delta_{1,\text{'quali3'}} = 0,\\ H_{0,4} : \delta_{2,\text{'quali1'}} = 0, \;\; H_{0,5} : \delta_{2,\text{'quali2'}} = 0, \;\; H_{0,6} : \delta_{2,\text{'quali3'}} = 0\}\end{aligned} \tag{2.13}$$

Note that there are no incremental effects here as in the case of duration dependence in treatment effects. This means that it does not make economic sense if any of these qualification differentiated treatment effects is negative. Hence, the family of alternative null hypotheses is one-sided

$$\begin{aligned}\{H_{A,1} : \delta_{1,\text{'quali1'}} > 0, \;\; H_{A,2} : \delta_{1,\text{'quali2'}} > 0, \;\; H_{A,3} : \delta_{1,\text{'quali3'}} > 0,\\ H_{A,4} : \delta_{2,\text{'quali1'}} > 0, \;\; H_{A,5} : \delta_{2,\text{'quali2'}} > 0, \;\; H_{A,6} : \delta_{2,\text{'quali3'}} > 0\}.\end{aligned} \tag{2.14}$$

Testing Results

The results from these qualification differentiated treatment effects (2.13) are listed in Table 2.4.

Each coefficient Coeff. in the first column is followed by its estimate $\widehat{\text{Coeff.}}$, and its t-statistic labeled as t. The columns labeled as individual tests list the critical values $c^{\cdot}$, p-values $p^{\cdot}$, and rejection at the 5% level rej$^{\cdot}$ of the individual tests. First under the normal distribution $c^{\mathcal{N}}_{0.95}$, $p^{\mathcal{N}}$, rej$^{\mathcal{N}}$, and second under bootstrapping $c^{boot}_{0.95}$, p^{boot}, rej^{boot}.

The columns labeled as 'multiple test' contain the multiple critical value $c^{MTest}_{0.95}$ such that the multiple error type one 'probability of falsely rejecting one or more null hypotheses' is controlled at the 5% level. This means that I control the risk of labeling one or more nonexistent treatment effects as statistically significant (one or more false discoveries) at the 5% level. The last column rej^{MTest} indicates which individual null hypotheses can be rejected using the multiple test instead of the individual tests.

$\delta_{1,\text{'quali3'}}$, $\delta_{2,\text{'quali3'}}$, and $\delta_{1,\text{'quali2'}}$ are found to be significantly larger than zero under individual testing with an $\mathcal{N}$ assumption. Under individual testing using bootstrapping instead of the $\mathcal{N}$ assumption, one also labels three treatment effect as statistically significant. However, these three individually significant treatment effects from individual testing are probably false discoveries, as multiple testing at the 5% significance level indicates. Under multiple testing, none of the six treatment effects are found to be statistically significantly greater than zero[9]. Thus, individual tests seem to falsely label these treatment effects as statistically significant at the 5% level, while multiple testing indicates that these are false discoveries at the 5% level.

Quantifying the Risk of Making False Discoveries by Individual Testing

Here again, the question is of what magnitude the risk of making false discoveries is by testing only individually. Note that the multiple critical value 4.43 puts a 5% threshold on the fam-

[9]Nonetheless, by controlling the FWE at the 10% level rather than at the 5% level, the three treatment effects that are significant under individual bootstrap testing remain statistically significant under multiple testing.

			individual tests at 5% level						multiple test 5% level	
Coeff.	$\widehat{\text{Coeff.}}$	t	$c_{0.95}^{\mathcal{N}}$	$p^{\mathcal{N}}$	rej$^{\mathcal{N}}$	$c_{0.95}^{boot}$	p^{boot}	rej^{boot}	$c_{0.95}^{MTest}$	rej^{MTest}
$\delta_{1,\text{'quali3'}}$	0.3282	4.39	1.64	0.000	yes	0.783	0.000	yes	4.43	no
$\delta_{2,\text{'quali3'}}$	0.1805	1.70	1.64	0.044	yes	1.233	0.019	no	4.43	no
$\delta_{1,\text{'quali2'}}$	0.3795	4.28	1.64	0.000	yes	0.687	0.002	yes	4.43	no
$\delta_{2,\text{'quali2'}}$	0.1855	1.40	1.64	0.081	no	1.344	0.044	yes	4.43	no
$\delta_{1,\text{'quali1'}}$	0.0409	0.80	1.64	0.211	no	4.371	0.953	no	4.43	no
$\delta_{2,\text{'quali1'}}$	0.0020	0.02	1.64	0.490	no	2.954	0.910	no	4.43	no

Table 2.4: Results from testing six treatment effects under one-sided alternatives

ilywise error rate[10]. The individual critical value under $\mathcal{N}$ at 1.64 is very far off the multiple critical value 4.43.

It is thus not surprising that the risk of making one or more false discoveries is around an unacceptably high 90% if one tests individually under the $\mathcal{N}$ distribution. The same note as in Subsection 2.5.1 concerning the two sources of the risk of making false discoveries applies here.

2.6 Conclusions

Is is important to test multiple effects in a multiple testing manner to guard against the danger of making false discoveries. If we test coefficients only individually by looking at their t-statistics, we run the danger of so-called false discoveries. That is, we run the danger of labeling treatment effects as statistically significant that do not exist. The simulation study illustrated that the higher the number of coefficients is that one looks at jointly, the higher is the risk of making such false discoveries. Furthermore, the lower the correlation is between the random variables associated with the null hypotheses, the higher is the risk of making one or more false discovery.

To this end, the well-known F-test or the more general Wald, Lagrange Multiplier, or Likelihood Ratio test have one major shortcoming. Any of these joint tests can essentially test just one **joint** null hypothesis against its **joint** alternative hypothesis. Thus, any of these tests can only tell us in special cases which individual coefficients contained in the joint null hypothesis are significant from a joint point of view; multiple testing methods generally allow for individual rejections.

The study from Lalive et al. (2005) that I reexamine uses a data set of 10,404 independent observations. Individual testing at the 5% significance level under the normal distribution translates into high risks of making false discoveries. Namely, the probabilities of falsely

[10]I.e., making one or more false discoveries

labeling one or more nonexistent treatment effects as statistically significant is around 20%[11] or even around 90%[12].

Multiple testing methods allow putting multiple treatment effects under joint statistical scrutiny, while controlling the risk of making false discoveries. Lalive et al. (2005) do not seem to have reported false discoveries, despite the fact that they just tested their treatment effects individually.

However, by differentiating treatment effects of benefit sanctions on the basis of qualifications[12], I provide evidence that individual t-tests probably make three false discoveries. That is, individual testing labels three treatment effects out of six as statistically significant. By putting a 5% threshold on the risk of making one or more false discoveries, these three individually significant treatment effects are indicated as false discoveries by multiple testing methods from Romano and Wolf (2005).

Unfortunately, most applied work does not take the risk of false discoveries into account, since only individual t-statistics are reported, or a test of one joint null hypothesis at the most. Among others, this paper and Heckman et al. (2010) highlight the need to control the risk of making false discoveries. From a meta point of view, the problem of false discoveries is even aggravated. If scientists only report results that work out of an actually much larger pool of candidate results they have tried, which Heckman et al. (2010) refer to as cherry picking, the danger of selectively reporting convenient results that may in fact be false discoveries undermines scientific credibility.

References

Please find the references of this essay at the end of this Ph.D. thesis.

[11] Four treatment effect coefficients, hence four null hypotheses
[12] Six treatment effect coefficients, hence six null hypotheses

2.A The Model that I Reexamine with Respect to False Discoveries

Let the random variable T_u denote the duration spent in unemployment. T_{s1} denotes the duration from entry to unemployment until a person gets a warning. Let T_{s2} denote the duration from a warning until the 100% benefit sanction is enforced, which is the case under Swiss law. The corresponding rates (2.15), (2.16), (2.17) of T_u, T_{s1}, T_{s2} are parameterized with observables only or with unobservables to allow for unobserved heterogeneity over individuals. In the first model without unobservables, Lalive et al. (2005) assume that the three rates of T_u, T_{s1}, T_{s2} are explained perfectly well by a set of observable variables. To make the model more realistic, they add unobserved heterogeneity terms u, v_1, v_2 within the rates θ_u, θ_{s1}, θ_{s2} as follows

$$\theta_u(t \mid x, D_1, D_2, u) = \lambda_u(t) \exp\{x'\beta_u + \delta_1 D_1 + \delta_2 D_2 + u\}, \tag{2.15}$$

$$\theta_{s_1}(t \mid x, v_1) = \lambda_{s_1}(t) \exp\{x'\beta_{s_1} + v_1\}, \tag{2.16}$$

$$\theta_{s_2}(t \mid x, v_2) = \lambda_{s_2}(t) \exp\{x'\beta_{s_2} + v_2\}. \tag{2.17}$$

which corresponds to equations (11) and (12) in Lalive et al. (2005).

(2.15) is the rate at which individuals drop out of employment. The higher θ_u is, the more likely is the favorable case that the individual drops out of unemployment. (2.16) is the rate of getting a warning. (2.17) is the rate of getting an unemployment benefit sanction. x denotes individual characteristics. Lalive et al. (2005) have an additional set of regressors k, which are public employment dummy variables. We could not get the data of these dummy variables k, thus our results do not perfectly match theirs. $D_1 = I(t_{s1} < t_u)$ is a dummy variable indicating if there was a warning for the individual. $D_2 = I(t_{s2} < t_u)$ denotes the dummy variable indicating if a sanction was enforced for the individual. The coefficients δ_1 and δ_2 are the two so-called ex-post treatment effects of benefit sanctions. Due to our missing public employment service dummy variables, we cannot estimate ex-ante treatment effects as in Lalive et al. (2005), which is a key innovation of their paper.

The $\lambda_\bullet(t)$ are coefficients, modeling flexible duration dependence with a step function. Generally speaking, the longer a person is unemployed, the less likely finding a job becomes, thus $\lambda_u(t)$ models negative duration dependence.

$$\lambda_u(t) = \exp\{\sum_{k=0}^{4} \lambda_{u,k} I_k(t)\}, \quad \lambda_{s1}(t) = \exp\{\sum_{k=0}^{4} \lambda_{s1,k} I_k(t)\}, \quad \lambda_{s2}(t) = \exp\{\sum_{k=0}^{4} \lambda_{s2,k} I_k(t)\} \tag{2.18}$$

The indicator functions $I_{\bullet,k}(t)$ are set up for the following time intervals, respectively: 0 to 3 months, 3 to 6 months, 6 to 9 months, 9 to 12 months, 12 and more months. Each $\lambda_{\bullet,0}$ is set to zero because a constant term is also estimated.

The model comprises a competing risks specification. That is, while a person is unemployed, he runs the competing risks of getting a benefit sanction warning or finding a job. Once the

person has got a warning, he runs the competing risks of a benefit sanction or finding a job. Once the person incurred a benefit sanction, he is left with the single risk of finding a job.

There are four treatment effect coefficients in Section 2.5.1 of this paper, which enter the three rates as follows

$$\begin{aligned}\theta_u(t \mid x, D_1, D_2, u) = \lambda_u(t) \exp\{x'\beta_u + \delta_{1,\text{0-29d}} D_{1,\text{0-29d}} + \delta_{1,\geq\text{30d}} D_{1,\geq\text{30d}} \\ + \delta_{2,\text{0-59d}} D_{2,\text{0-59d}} + \delta_{2,\geq\text{60d}} D_{2,\geq\text{60d}} + u\}, \qquad (2.19)\end{aligned}$$

$$\theta_{s_1}(t \mid x, v_1) = \lambda_{s_1}(t) \exp\{x'\beta_{s_1} + v_1\}, \qquad (2.20)$$

$$\theta_{s_2}(t \mid x, v_2) = \lambda_{s_2}(t) \exp\{x'\beta_{s_2} + v_2\}. \qquad (2.21)$$

In Section 2.5.2 of this paper, I differentiated treatment effects based on three levels of qualifications 'quali1', 'quali2', 'quali3', which are abbreviated as $q1$, $q2$, $q3$ here, respectively. Thus, there are three rates $\theta_{u,q1}$, $\theta_{u,q2}$, $\theta_{u,q3}$ instead of just one rate θ_u as before. The resulting five rates are

$$\theta_{u,q1}(t \mid x, D_1, D_2, u_{q1}) = \lambda_{u,q1}(t) \exp\{x'\beta_{u,q1} + \delta_{1,q1} D_1 + \delta_{2,q1} D_2 + u_{q1}\}, \qquad (2.22)$$

$$\theta_{u,q2}(t \mid x, D_1, D_2, u_{q2}) = \lambda_{u,q2}(t) \exp\{x'\beta_{u,q2} + \delta_{1,q2} D_1 + \delta_{2,q2} D_2 + u_{q2}\}, \qquad (2.23)$$

$$\theta_{u,q3}(t \mid x, D_1, D_2, u_{q3}) = \lambda_{u,q3}(t) \exp\{x'\beta_{u,q3} + \delta_{1,q3} D_1 + \delta_{2,q3} D_2 + u_{q3}\}, \qquad (2.24)$$

$$\theta_{s_1}(t \mid x, v_1) = \lambda_{s_1}(t) \exp\{x'\beta_{s_1} + v_1\}, \qquad (2.25)$$

$$\theta_{s_2}(t \mid x, v_2) = \lambda_{s_2}(t) \exp\{x'\beta_{s_2} + v_2\}. \qquad (2.26)$$

Lalive et al. (2005) estimate the model by maximizing the resulting closed-form log-likelihood function. The Heckman-Singer mass point approach is used to estimate the model with unobservables by maximum likelihood, as in Heckman and Singer (1984). I used TSP (Time Series Package) to estimate the model, as Lalive et al. (2005) did. The ML solver of TSP can make use of analytic derivatives, which is a neat feature for the closed-form densities of the model. These authors did not bootstrap the model, however, they rely on asymptotic normality to judge the significance of the estimated treatment effects.

2.B Implementation

The short code of the simulation study is available from the author on request; I do not elaborate on it here. Nonetheless, I elaborate on the implementation of the empirical part of this paper, which consisted of the following five steps

1. Replicate Lalive et al. (2005)'s results
2. Implement four and six treatment effects model based on replicating code
3. Bootstrap the four and six treatment effects model

4. Do individual and multiple testing of four or six treatment effects based on bootstrap results

5. Quantify the risk of making one of more false discoveries by individual instead of multiple testing

2.B.1 Replicate their results

I got the original TSP code (Time Series Package) for the basic two treatment effects model in Lalive et al. (2005) from Raphael Lalive, to which I could compare my implementation. He also provided me with the original data, except for the public employment dummies, which he was not allowed to pass on to me due to data protection issues.

2.B.2 Implement Four and Six Treatment Effects Model Based on Replicating Code

Rafael Lalive also helped me set up the code for the four treatment effects model that is contained in Lalive et al. (2005). Based on these two codes, I coded the six treatment effects model. Details of the six treatment effects model can be found in Appendix 2.A.

2.B.3 Bootstrap the Four and Six Treatment Effects Model

Bootstrapping my two models was fairly easy, given that one can use the conventional i.i.d. bootstrap by drawing single data points with replacement from the original data.

Specifically, let $\hat{\theta}$ denote the ML coefficients of one of my two models using the original data

$$\hat{\theta} = \arg\max_{\theta} L(\theta; data), \tag{2.27}$$

where $data$ denotes the $[10404 \times p]$ matrix containing the original data and $L(\cdot;\cdot)$ denotes the (log) likelihood. Given no dependence between individuals but possible contemporaneous dependence between variables, one can generate an artificial bootstrap data set $data_1^*$ by drawing single data rows with replacement $10,404$ times from the sequence of data rows $1, 2, \ldots, 10404$. Note that by falsely i.i.d bootstrapping each variable, that is column, separately, the possible contemporaneous dependence of the variables is destroyed. Thus, the first bootstrap data set of size $[10404 \times p]$ may be

$$\text{First bootstrap data set } data_1^* : \underbrace{\text{data row}_{235}, \text{data row}_{52}, \text{data row}_{9874}, \ldots, \text{data row}_{52}}_{\text{10,404 data rows as in original data set}}$$

The second bootstrap data set of size $[10404 \times p]$ may look like

$$\text{Second bootstrap data set } data_2^* : \underbrace{\text{data row}_{189}, \text{data row}_{8532}, \text{data row}_{10203}, \ldots, \text{data row}_{9874}}_{\text{10,404 data rows as in original data set}}$$

In this way, I generated 2,500 bootstrap data sets $data_1^*$, $data_2^*$, ..., $data_{2500}^*$. By computing the ML coefficients on each of these 2,500 bootstrap data sets, I get 2,500 bootstrap ML coefficients

$$\begin{aligned} \hat{\theta}_1^* &= \arg\max_\theta L(\theta; data_1^*), \\ \hat{\theta}_2^* &= \arg\max_\theta L(\theta; data_2^*), \\ &\vdots \\ \hat{\theta}_{2500}^* &= \arg\max_\theta L(\theta; data_{2500}^*). \end{aligned} \tag{2.28}$$

2.B.4 Do Individual and Multiple Testing of Four or Six Treatment Effects Based on Bootstrap Results

Individual One- and Two-Sided Bootstrap Tests

To carry out an individual t-test of an individual coefficient $\beta \in \theta$ based on bootstrapping instead of an $\mathcal{N}$-assumption, one computes the t-value of the individual coefficient β on each of these bootstrap data sets, resulting in 2,500 t-values $\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*)$, $\hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*)$, ..., $\hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*)$[13].

The bootstrap critical value at significance level α for a one-sided bootstrap test with alternative $H_A : \beta > 0$ is just the $1-\alpha$ empirical quantile of the 2,500 t-values $\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*)$, $\hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*)$, ..., $\hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*)$.

The bootstrap critical value at significance level α for a two-sided bootstrap test with alternative $H_A : \beta \neq 0$ is the $1-\alpha$ empirical quantile of the 2,500 t-values in absolute value $\left|\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*)\right|$, $\left|\hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*)\right|$, ..., $\left|\hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*)\right|$.

If the observed t-value $\hat{\beta}/\hat{\sigma}(\hat{\beta})$ is larger than the $1-\alpha$ bootstrap critical value based on the 2,500 bootstrap data sets, the null hypothesis is rejected at significance level α.

The one-sided bootstrap p-value for coefficient $\beta \in \theta$ is computed as

$$p^{boot} = \frac{\sum_{m=1}^{2500} \mathbf{1}\{\hat{\beta}_m^*/\hat{\sigma}^*(\hat{\beta}_m^*) > \hat{\beta}/\hat{\sigma}(\hat{\beta})\}}{2500+1}, \tag{2.29}$$

where $\mathbf{1}$ denotes the indicator function. In the two-sided case, the absolute values of the bootstrap and the original t-values must be used to compute the bootstrap p-value.

[13]The standard deviation $\hat{\sigma}(\hat{\beta}_m^*)$ on bootstrap data set data_m^* was computed by the Eicker-White method, which is a combination of analytic second derivatives and the covariance of the analytic gradient. Asymptotically, these two ways of computing the standard deviation of an ML estimator obtain the same result, as stipulated in the so-called information matrix equality. On the computer, these two ways may yield different results, though. The Eicker-White estimator finds an optimal combination thereof. This corresponds to the `HCOV=W` option in TSP's `ML()` routine.

Multiple Testing

The implementation of the multiple test is straight forward. On Michael Wolf's webpage, there is R and Matlab code available that carries out multiple testing. One can simply pass the vector of observed t-statistics $[\hat{\beta}/\hat{\sigma}(\hat{\beta}), \beta \in \theta]$ and the matrix of bootstrap t-statistics $[\hat{\beta}^*_1/\hat{\sigma}(\hat{\beta}^*_1), \hat{\beta}^*_2/\hat{\sigma}(\hat{\beta}^*_2), \ldots, \hat{\beta}^*_{2500}/\hat{\sigma}(\hat{\beta}^*_{2500}), \beta \in \theta]$ to the R or Matlab code. The bootstrap matrix $[\hat{\beta}^*_1/\hat{\sigma}(\hat{\beta}^*_1), \hat{\beta}^*_2/\hat{\sigma}(\hat{\beta}^*_2), \ldots, \hat{\beta}^*_{2500}/\hat{\sigma}(\hat{\beta}^*_{2500}), \beta \in \theta]$ consists of 2,500 rows, each row containing the estimated t-statistic for the m^{th} bootstrap data set. As said before, for the two-sided alternative hypothesis case, element-wise absolute values must be provided.

Thus, for my four treatment effects model, I passed a $[4 \times 1]$ vector of observed t-statistics and a $[2500 \times 4]$ matrix of bootstrap t-statistics to the R function, since $[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4] = \hat{\theta}$. For the six treatment effects model, I passed a $[6 \times 1]$ vector of observed t-statistics and a $[2500 \times 6]$ matrix of bootstrap t-statistics to the R function.

To control the familywise error rate at the 5% level, one must additionally provide the R or Matlab function with $k = 1$ and $\alpha = 0.05$. The R or Matlab function then returns the multiple critical value and the indices of the null hypotheses that could be rejected while controlling the familywise error rate at the α% level.

2.B.5 Quantify the Risk of Making One of More False Discoveries by Testing Individually Instead of Multiple Testing

Quantifying the risk of making one or more false discoveries by assuming $\mathcal{N}$ instead of relying on multiple testing is easy. The multiple testing method of Romano and Wolf (2005) based on bootstrapping essentially solves the following problem

MTest: Find one critical value $c^{MTest}_{1-\alpha}$ such that familywise error rate FWE $\leq \alpha$ asymptotically.

Quantifying the probability of making one or more false discoveries $\widehat{\text{FWE}}$ in assuming $\mathcal{N}$ essentially solves the related problem

Find $\widehat{\text{FWE}}$: Find $\hat{\alpha}$ such that the $\mathcal{N}$-assumed critical value $c^{\mathcal{N}\text{ass.}}_{1-\alpha}$ ensures FWE $\leq \hat{\alpha}$ asymptotically.

This can be done by a grid search over the parameter α that is passed on to the multiple testing routine. For example, in the two-sided testing, the $\mathcal{N}$ assumption results in the critical value 1.96 at the 5% level. I quantify the committed multiple error type one $\widehat{\text{FWE}}$ in using 1.96 instead of the multiple critical value as follows. Increase $\hat{\alpha}$ passed to the multiple testing routine until a multiple critical value of 1.96 is returned. The $\hat{\alpha}$ that satisfies this criterion up to two decimal points is the approximate risk of making one or more false discoveries $\widehat{\text{FWE}}$ in assuming $\mathcal{N}$ instead of relying on multiple testing.

Note: Two sources of error in individually testing under normal assumption It would be very interesting to know exactly whether the high risk of making one or more false

discoveries by assuming $\mathcal{N}$ is mostly due to individual bootstrap testing instead of multiple bootstrap testing. Or whether it is mainly due to assuming $\mathcal{N}$ instead of bootstrapping the data generating distribution. I expect the error stemming from assuming $\mathcal{N}$ instead of bootstrapping to be small, as a central limit theorem suggests for $10,404$ data points. Nonetheless, to answer this question exactly, the easy approximation scheme as in step (e) does not work for this case, since there is not a single individual bootstrap critical value covering all null hypotheses. For example, the individual two-sided bootstrap critical values $c^{boot}_{|\cdot|,0.95}$ range from 1.97 to 2.17 for the four treatment effects model.

Hence, one needs a two-stage bootstrap analysis to answer this question, which is very computing intensive for this nonlinear ML problem. Specifically, one needs not only carry out an ML routine on each of the 2,500 bootstrap data sets, as was the case to carry out multiple testing. One even needs to conduct a so-called second-stage bootstrap analysis of 500 repetitions, say, on each of the 2,500 first-stage bootstrap samples, to answer this question. Hence, the ML problem needs to be solved $2,500 \times 501$ times, which takes a long time of parallel computing.

2.C Control of More Liberal Multiple Error Types

Four treatment effects: Duration Dependence in Treatment Effects Note that I control the 'probability of falsely rejecting one or more null hypotheses' at the 5% level.

What happens if we get more liberal with respect to false discoveries, thus merely want to control the 'probability of falsely rejecting **two or more** null hypotheses'? What if we even considered the very liberal 'probability of falsely rejecting **four or more** null hypotheses'?

What essentially happens is that the more liberal the multiple error type one gets in the aforementioned sense, the lower the critical value gets that the multiple testing method returns. Controlling the 'probability of falsely rejecting **two or more** null hypotheses' at the 5% level is achieved by a multiple critical value of 1.36. The two multiple critical values 1.00 and 0.80 control the multiple error types I '**three or more ...**' and '**four or more ...** ' at the 5% level, respectively. Thus, the higher one sets the number of false discoveries in the risk threshold, the lower the critical value gets, thus the more treatment effects are labeled as statistically significant. By merely individually testing, one knows that the risk of false discoveries is present, but one cannot put a risk threshold on it. Unfortunately, this seems to be the modus operandi in most applied work.

One can avoid choosing the k in controlling the 'probability of falsely rejecting k **or more** null hypotheses' by considering **relative** multiple error types I, such as the False Discovery Proportion (FDP). Control of the FDP is achieved by increasing k within 'probability of falsely rejecting k or more null hypotheses' until a criterion is met; see Romano and Wolf (2005) for details.

Six treatment effects: Qualification Dependence in Treatment Effects The more liberal probability of falsely declaring two or more false discoveries is still at 33.3% by the individual critical value 1.64 resulting from assuming $\mathcal{N}$. Recall that the probability of falsely rejecting one or more null hypotheses was around 90%.

The $\mathcal{N}$ derived individual critical value 1.64 puts a 1.9% threshold on the very relaxed multiple error type one 'probability of falsely declaring three or more treatment effects as statistically significant'. Thus, it is perfectly possible that by individually testing, one does control the risk of making a number of false discoveries at conventional significance levels by coincidence. But again, multiple testing methods let us quantify and put a threshold on the risk of making false discoveries.

Essay 3

Joint Prediction Regions

Bootstrap Joint Prediction Regions

Michael Wolf[1]
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

Dan Wunderli[2]
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
dan.wunderli@econ.uzh.ch

January 2012

Abstract

Many economic and financial applications require the forecast of a random variable of interest over several periods into the future. The sequence of individual forecasts, one period at a time, is called a path forecast, where the term path refers to the sequence of individual future realizations of the random variable, one period at a time. The problem of constructing a corresponding joint prediction region has been rather neglected in the literature so far: such a region is supposed to contain the entire future path with a prespecified probability. We develop bootstrap methods to construct joint prediction regions. The resulting regions are proven to be asymptotically consistent under a mild high-level assumption. We compare the finite-sample performance of our joint prediction regions to some previous proposals via Monte Carlo simulations. An application to real data is also provided.

An altered version of this paper is currently under review at the Journal of Time Series Analysis.

KEY WORDS: Bootstrap; generalized error rates; joint prediction regions.

JEL CLASSIFICATION NOS: C32, C53.

[1]Research has been supported by the NCCR Finrisk project "New Methods in Theoretical and Empirical Asset Pricing".

[2]Thanks to participants of the Zurich Workshop in Economics 2011, Hotel Seeburg, Lucerne.

3.1 Introduction

When predicting a random variable, a point forecast alone is often considered insufficient. In addition, a statement about the uncertainty contained in the point forecast, as expressed by a *prediction interval*, may also be desired.

This is similar to the situation where a point estimator of a population parameter alone is considered insufficient; and where a statement about the uncertainty contained in the point estimate, as expressed by a *confidence interval*, is also desired.

Constructing a prediction interval for a random variable is inherently more difficult than constructing a confidence interval for a population parameter.

In the latter problem, typically a central limit theorem can be applied to argue that an estimator of the parameter has, approximately, a normal distribution for large sample sizes. This allows for standard, normal-theory confidence intervals described in any basic statistics text book. The use of bootstrap methods as an alternative is 'only' motivated by higher-order considerations; standard methods already result in confidence intervals that are consistent, that is, have coverage probability equal to the nominal level $1 - \alpha$ asymptotically.

In the former problem, no central limit theorem can be applied to argue that the difference between a point forecast and the random variable of interest has, approximately, a normal distribution for large sample sizes; for example, such an assumption is made by Jordà and Marcellino (2010). Therefore, standard normal-theory prediction intervals are only valid, even asymptotically, under restrictive additional assumptions, such as a linear time series with normal errors. The use of bootstrap methods as an alternative is motivated by first-order considerations already; they result in prediction intervals that are consistent under very general assumptions where standard, normal-theory prediction intervals fail.

How to apply the bootstrap to construct prediction intervals that are not only asymptotically consistent but also have good finite-sample properties is not a trivial problem. But it can be considered solved by now to a satisfactory degree; for example, see De Gooijer and Hyndman (2006, Section 12) for an overview.

The discussion so far only applies to a single (future) random variable. In many applications, however, a random variable of interest is predicted up to H periods into the future. For example, one might predict future inflation for the next $H = 12$ months. A *path* refers to the sequence of future realizations 1 to H periods into the future. A *path-forecast* refers to the sequence of corresponding forecasts 1 to H periods into the future.

On the one hand, one can construct marginal prediction intervals by using a given method repeatedly to construct a prediction interval for a single future random variable, one period at a time. But, by design, probability statements then only apply marginally, one period at a time: the prediction interval at a specific horizon h, for some $1 \leq h \leq H$, will contain the random variable h periods into the future with a prespecified probability $1 - \alpha$.

On the other hand, a more general problem is the construction of a *joint prediction region* that will contain the entire path with the desired probability $1-\alpha$. For example, if one would like to know how high inflation might rise over the next $H = 24$ months, with probability $1-\alpha$, one needs to construct a joint prediction region for the future path at level $1-\alpha$ as opposed to stringing together 12 marginal prediction intervals, each one at level $1-\alpha$.

It should be clear that stringing together marginal prediction intervals for horizons $h = 1$ up to $h = H$, each one at level $1-\alpha$, will not result in a joint prediction region that contains the entire path with probability $1-\alpha$. Instead, apart from pathological cases, the joint coverage probability will be strictly less than $1-\alpha$, and decreasing in H. Denote by E_h the event that the random variable at h periods in the future will fall into its prediction interval. If the events E_h are independent of each other, then stringing together marginal prediction intervals results in a joint prediction region that will contain the entire path with probability $(1-\alpha)^H$ only.[3]

The construction of joint prediction regions for future paths of a random variable of interest has been rather neglected in the forecasting literature so far. Two notable exceptions are Jordà and Marcellino (2010) and Staszewska-Bystrova (2010). The former work proposes an 'asymptotic' method that relies on the overly strong assumption that forecast errors have, approximately, a normal distribution. The latter work proposes a bootstrap method that is of heuristic nature only. Therefore, neither of the proposed methods appears safe to use in practice.

In this paper, we propose a bootstrap method to construct joint predictions regions that are proven to contain future paths of random variable of interest with probability $1-\alpha$, at least asymptotically, under a mild high-level assumption.

In addition, we also consider the more general problem of constructing joint confidence regions that will only contain all elements of future paths up to a small number $k-1$ of them with probability $1-\alpha$. If the maximum forecast horizon H is large, the applied researcher may deem the criterion that all elements of the future path must be contained in the joint prediction region with probability $1-\alpha$ as too strict. For example, when $H = 24$, it may be deemed acceptable that up to $k-1 = 2$ elements of the future path may fall outside the joint prediction region; thus requiring that 'only' at least 22 of the 24 elements — or at least 90% of the 24 elements — of the future path be contained in the joint prediction region with probability $1-\alpha$. The choice of k must be made by the applied researcher, not by the statistician. But it will be useful to the applied researcher to have a method available that can handle any desired value of k. In particular, the choice $k = 1$ yields a 'standard' joint prediction region that must contain all elements of a future path with probability $1-\alpha$.

The remainder of the paper is organized as follows. Section 3.2 contains some background results that are useful for setting the stage. Section 3.3 describes our method to construct joint prediction regions and compares it to some previous proposals in the literature. Sec-

[3] In practice, the events E_h are typically not independent of each other. Stringing together marginal prediction intervals then results in a joint prediction region that will contain the entire path with probability somewhere between $(1-\alpha)^H$ and $1-\alpha$. The exact probability is a function of the dependence structure of the events E_h.

tion 3.4 studies finite-sample performance via Monte Carlo simulations. Section 3.5 provides an empirical application to real data. Finally, Section 3.6 concludes. All mathematical proofs and some further background results are collected in an appendix.

3.2 Background Results

Our motivating problem is the construction of a joint prediction region for a future path of a random variable of interest. However, the proposed methodology applies more generally to the construction of a joint prediction region of an arbitrary random vector that has not been observed yet.

In explaining the methodology, it will be convenient to start with the special case of a single random variable that has not been observed yet.

3.2.1 Single Forecast

First, consider a single random variable y with mean $\mu \equiv \mathbb{E}(y)$. This special case makes it easier to explain some fundamental concepts before considering the more general case of a random vector with H elements.

One may wish to predict y or to estimate μ. Denote the forecast of y by $\widehat{y}$ and the estimator of μ by $\widehat{\mu}$. Often times, the two are actually the same, that is $\widehat{y} = \widehat{\mu}$; for example, in the context of linear regression models. Therefore, in terms of a (point) forecast of y compared a (point) estimate of μ, there often is no difference at all.

But what if one desires an 'uncertainty interval' in addition? Such an interval should contain the random variable y or its mean μ, respectively, with a prespecified probability $1 - \alpha$. (To be careful, this probability only exists before computing the interval from a frequentist view point, at least for the mean μ.) Now the two solutions are fundamentally different and the former interval will have to be wider due to the additional randomness contained in the random variable y compared to its mean μ. To make this distinction apparent in the notation, we prefer to call the solution to the former problem a *prediction interval* and the solution to the latter problem a *confidence interval*. In doing so, we are in agreement with De Gooijer and Hyndman (2006, p.460):

> Unfortunately, there is still some confusion in terminology with many authors using "confidence interval" instead of "prediction interval". A confidence interval is for a model parameter, whereas a prediction interval is for a random variable. Almost always, forecasters will want prediction intervals — intervals which contain the true values of future observations with [a] specified probability.

It is also useful to point out that there is a duality between a confidence interval for μ and a

hypothesis test for μ. For concreteness, consider the two-sided hypothesis testing problem

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 \ . \tag{3.1}$$

If a level α test is available for this problem, for every value of μ_0, then a two-sided confidence interval CI for μ at level $1 - \alpha$ can be constructed by *inverting* the hypothesis test as follows:

$$\text{CI} \equiv \{\mu_0 : \mu_0 \text{ is not rejected by the hypothesis test}\} \ . \tag{3.2}$$

That is, the collection of values μ_0 not rejected by the test at significance level α constitutes a confidence interval with confidence level $1 - \alpha$. Conversely, a hypothesis test for problem (3.1) can be carried by *inverting* a two-sided confidence interval for μ: one simply rejects H_0 at significance level α if and only if (iff) the value μ_0 is not contained in the confidence interval with confidence level $1 - \alpha$.

Analogously, there is a duality between a one-sided confidence interval for μ and a one-sided hypothesis test for μ; the details are straightforward.

On the other hand, there is no duality between a prediction interval for y and a hypothesis test for y. This is because y is a random variable and not a (non-random) parameter and hypothesis tests on such random quantities do not exist. In particular, the testing problem

$$H_0 : \widehat{y} - y = 0 \quad \text{versus} \quad H_1 : \widehat{y} - y \neq 0 \tag{3.3}$$

is nonsensical. The quantity $\widehat{y} - y$ is a random variable. If its distribution is continuous, then $\widehat{y} - y$ will be different from zero with probability one, irrespective of the 'quality' of the forecast $\widehat{y}$.

3.2.2 Path-Forecast

More generally, consider a random vector $Y \equiv (y_1, \ldots, y_H)'$ of interest with mean $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_H)' = \mathbb{E}(Y)$. For the purposes of this paper, Y will typically correspond to the values of a random variable one to H periods into the future; that is, to a future *path* of a random variable. But the discussion below applies to any random vector. The underlying probability mechanism is denoted by $\mathbb{P}$.

One can wish to predict Y or to estimate $\boldsymbol{\mu}$. Denote the forecast of Y by $\widehat{Y}$ and the estimator of $\boldsymbol{\mu}$ by $\widehat{\boldsymbol{\mu}}$. (When Y corresponds to a future path of a random variable, $\widehat{Y}$ is also called a *path-forecast*.) Again, often times, the two are actually the same, that is, $\widehat{Y} = \widehat{\boldsymbol{\mu}}$; for example, in the context of linear regression models. Therefore, again, in terms of a (point) forecast of Y compared to a (point) estimate of $\boldsymbol{\mu}$, there often is no difference at all.

What if one desires the extension of an 'uncertainty interval' for a univariate quantity to a '(joint) uncertainty region' for a multivariate quantity? In the most stringent case, such a region should contain the *entire* random vector Y or its mean $\boldsymbol{\mu}$, respectively, with a prespecified probability $1 - \alpha$. Again, the two solutions are fundamentally different and the former region

will have to be larger (in volume) due to the additional randomness contained in Y compared to $\boldsymbol{\mu}$.

Again, there is a duality between a joint confidence region for the parameter $\boldsymbol{\mu}$ and a a hypothesis test for $\boldsymbol{\mu}$. In the multivariate setting, the testing problem is inherently of a two-sided nature:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \ . \tag{3.4}$$

If a level α test is available for this problem, for every value of $\boldsymbol{\mu}_0$, then a joint confidence region JCR for $\boldsymbol{\mu}$ at level $1-\alpha$ can be constructed by *inverting* the hypothesis test as follows:

$$\text{JCR} = \{\boldsymbol{\mu}_0 : \boldsymbol{\mu}_0 \text{ is not rejected by the hypothesis test}\} \ . \tag{3.5}$$

That is, the collection of values not rejected by the test at significance level α constitutes a joint confidence region with confidence level $1-\alpha$. Conversely, a hypothesis test for problem (3.4) can be carried out by *inverting* a joint confidence region for $\boldsymbol{\mu}$: one simply rejects H_0 at significance level α iff the value $\boldsymbol{\mu}_0$ is not contained in the joint confidence region with confidence level $1-\alpha$.

Again, on the other hand, there is no duality between a joint prediction region for Y and a hypothesis test for Y. The latter does not exist because Y is a random vector and not a (non-random) parameter.

A potential complication with joint regions arises when uncertainty statements concerning the individual components y_h or μ_h, respectively, are desired. For example, this is typically the case when a joint prediction region for Y is to be constructed in addition to a path-forecast $\widehat{Y}$. One desires lower and upper bounds for each component y_h in such a manner that the entire vector Y be contained in the implied rectangle with probability $1-\alpha$. This is a trivial task if the underlying joint prediction region is already of rectangular form. But this is not true for all methods to compute joint regions; many methods result in regions of elliptical form instead. The most prominent example is the Scheffé joint region, dating back to Scheffé (1953); Scheffé (1959).

The Scheffé joint confidence region for $\boldsymbol{\mu}$ is obtained by inverting the classical F-test. Let $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\mu}})$ denote an estimated covariance matrix of $\widehat{\boldsymbol{\mu}}$. Then the joint confidence region is given by

$$\text{JCR} \equiv \left\{\boldsymbol{\mu}_0 : (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)' \left[\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\mu}})\right]^{-1} (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \leq \chi^2_{H,1-\alpha}\right\} \ , \tag{3.6}$$

where $\chi^2_{H,1-\alpha}$ denotes the $1-\alpha$ quantile of the chi-square distribution with H degrees of freedom. The use of this joint confidence region is usually justified by a central limit theorem implying an approximate multivariate normal distribution of $\widehat{\boldsymbol{\mu}}$ with mean $\boldsymbol{\mu}$. Such a central limit theorem will hold under mild regularity conditions; for example, see White (2001).

The Scheffé joint prediction region for Y is obtained similarly. Define the vector of prediction errors by $\widehat{U} \equiv \widehat{Y} - Y$ and let $\widehat{\boldsymbol{\Sigma}}(\widehat{U})$ denote an estimated covariance matrix of this vector. Then

the joint prediction region is given by

$$\text{JPR} \equiv \left\{ X : (\widehat{Y} - X)' \left[\widehat{\boldsymbol{\Sigma}}(\widehat{Y}) \right]^{-1} (\widehat{Y} - X) \leq \chi^2_{H,1-\alpha} \right\} . \tag{3.7}$$

The use of this joint prediction region is only justified if $\widehat{U}$ has approximately a multivariate normal distribution with mean zero. This is a strong additional assumption, which is often violated in practice. A central limit theorem can typically be applied to argue that an estimator has, approximately, a normal distribution for large sample sizes. But a central limit theorem can never be applied to argue that a forecast error has, approximately, a normal distribution for large sample sizes. This point is illustrated via a simple example in Remark 3.3.2 below.

If the joint region is of elliptical form and statements concerning the individual components are desired, the joint region has to be 'projected' on the axes of $\mathbf{R}^H$. This action implies a *larger* rectangular joint region: namely, the smallest rectangle, with sides parallel to the axes of $\mathbf{R}^H$, that contains the original elliptical region. As a result, if the elliptical region has joint coverage probability of $1-\alpha$, then the implied rectangular region has joint coverage probability larger than $1 - \alpha$. Therefore, such a projection method is generally overly conservative. If statements concerning the individual components are desired, it is better to construct 'direct' rectangular joint regions instead; that is, joint regions that are designed to be of rectangular form to begin with.

Remark 3.2.1. It will be useful to illustrate these concepts in simple, parametric setup. Assume $Y \equiv (Y_1, Y_2)' \sim N(\boldsymbol{\mu}, \boldsymbol{I}_2)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{I}_2$ is the identity matrix of dimension two. Therefore, Y_1 and Y_2 are independent with $Y_h \sim N(\mu_h, 1)$. The goal is to construct a joint confidence region for $\boldsymbol{\mu}$. The point estimator for $\boldsymbol{\mu}$ is simply given by the observed random vector, that is, $\widehat{\boldsymbol{\mu}} \equiv Y$.

The Scheffé joint confidence region is obtained by inverting the classical F-test. It is a circle centered at Y with radius $\sqrt{\chi^2_{2,1-\alpha}}$, where $\chi^2_{2,1-\alpha}$ denotes the $1-\alpha$ quantile of the chi-square distribution with two degrees of freedom. For example, when $\alpha = 0.05$, the radius is $\sqrt{5.99} = 2.45$. The implied rectangular joint confidence region, obtained by projecting the circle on the two axes, is a square with center Y and half length 2.45.

On the other hand, a 'direct' rectangular joint confidence region is given by

$$\left[Y_1 \pm d_{2,1-\alpha} \right] \times \left[Y_1 \pm d_{2,1-\alpha} \right] ,$$

where $d_{2,1-\alpha}$ is the $1 - \alpha$ quantile of the random variable $\max\{|Y_1 - \mu_1|, |Y_2 - \mu_2|\}$. These quantiles are not commonly tabulated, but can be easily simulated to arbitrary precision. For example, when $\alpha = 0.05$, then $d_{2,0.95} = 2.24$.

The 'direct' rectangular joint confidence region is thus a square with center Y and half length 2.24. Therefore, it is smaller than the implied rectangular joint confidence region by the Scheffé method.

The Scheffé region itself has a smaller volume than the 'direct' rectangular region when $\alpha =$

0.05, namely

$$2.45^2 \cdot \pi = 18.86 < 20.07 = (2 \cdot 2.24)^2 \ .$$

But when a rectangular region is needed in the end, projecting the Scheffé region on the axes results in a larger region compared to the 'direct' rectangular region. An illustration is provided in Figure 3.1. ■

The stringent joint regions discussed so far control the probability of containing the entire vector of interest to be (at least) equal to $1 - \alpha$. Equivalently, they control the probability of missing at least one component of the vector to be (at most) equal to α. Borrowing from the multiple testing literature, the latter probability can be termed the *familywise error rate* (FWE); for example, see Romano et al. (2008). So for a joint confidence region (JCR) for $\boldsymbol{\mu}$,

$$\text{FWE} \equiv \mathbb{P}\{\text{At least one of the } \mu_h \text{ not contained in the JCR}\} \ , \tag{3.8}$$

whereas for a joint prediction region (JPR) for Y,

$$\text{FWE} \equiv \mathbb{P}\{\text{At least one of the } y_h \text{ not contained in the JPR}\} \ . \tag{3.9}$$

Jordà et al. (2010, Section 2.2) argue that controlling the FWE can be too strict:

> For example, in a prediction of a path of monthly inflation over the next two years, control of the FWE would result in rejection of such paths as when the trajectory of inflation is [almost] correctly predicted for 23 periods but the prediction of the last month is particularly poor.

The decision whether the FWE is too strict or not in a given application has to be made by the applied researcher, not by the econometrician. It is the job of the econometrician to provide the applied researcher with an alternative tool in case his decision is against control of the FWE. Jordà et al. (2010) propose as an alternative to control the *false discovery rate* (FDR). Unfortunately, this proposal is actually equivalent to control of the familywise error rate in the context of joint confidence regions and joint prediction regions. An explanation of this fact is a bit lengthy and can be found in Appendix 3.B.

While control of the FDR is not a meaningful alternative, it is possible to construct joint confidence regions as well as joint prediction regions based on a generalized error rate that is meaningful in the context of joint regions. The solution is to use the *generalized familywise error rate* (k-FWE).

For a joint confidence region (JCR) for $\boldsymbol{\mu}$,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } \mu_h \text{ not contained in the JCR}\} \ , \tag{3.10}$$

whereas for a joint prediction region (JPR) for Y,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } y_h \text{ not contained in the JPR}\} \ . \tag{3.11}$$

As a special case, the choice $k = 1$ gives back the FWE. On the other hand, any choice $k \geq 2$ results in a less stringent error rate.

As will be discussed in Section 3.3, the larger the value of k the smaller the resulting joint region. Consequently, by being willing to miss a small number of components in the joint region, the applied researcher can obtain more precise bounds in return.

Since the number of components, H, is known, control of the k-FWE immediately gives control on the probability of the proportion of components not contained in the joint region. Take the example of a path-forecast with $H = 24$ components, as when predicting monthly inflation for the next two years. Then the choice $k = 3$ allows for a proportion of missed components up to 10%. This is because one or two missed components, out of the $H = 24$, do not constitute a violation of the event underlying the error rate, but three or more missed components do.

The next section details how the k-FWE, which includes the FWE as a special case, can be controlled in practice. It only does this in the context of a joint prediction region for Y. The method is analogous in the context of a joint confidence region for $\boldsymbol{\mu}$ and is detailed in Romano and Wolf (2007) already.

Because the method is based on quantiles of random variables whose cumulative distribution function may not be invertible, the following remark is in order.

Remark 3.2.2. If the cumulative distribution function of a random variable is not invertible, then its quantiles are not necessarily uniquely defined. For concreteness, we adopt the following definition for quantiles in this paper.

Let X be a random variable with cumulative distribution function $F(\cdot)$. Then, for $\lambda \in (0, 1)$, the λ quantile of (the distribution) of X is defined as $\inf\{x : F(x) \geq \lambda\}$. ■

3.3 Joint Prediction Regions Based on k-FWE Control

The goal is to construct a joint prediction region for Y that controls the k-FWE, for an arbitrary integer $1 \leq k < H$. In particular, the special choice $k = 1$ corresponds to control of the FWE.

Any formal analysis has to be put into a suitable framework. To this end, we borrow some notation from Jordà et al. (2010). We start out by discussing the case of a univariate time series, which simplifies the notation and makes it easier to focus on the methodology.

3.3.1 Univariate Time Series

One observes a univariate time series $\{y_1, \ldots, y_T\}$ generated from a true probability mechanism $\mathbb{P}$ and wishes to predict the next stretch of H observations $\{y_{T+1}, \ldots, y_{T+H}\}$. Let $Y_{T,H} \equiv (y_{T+1}, \ldots, y_{T+H})'$. At time t, denote a prediction h periods ahead by $\widehat{y}_t(h)$. Then a

path-forecast for $Y_{T,H}$ is given by $\widehat{Y}_T(H) \equiv (\widehat{y}_T(1), \ldots, \widehat{y}_T(H))'$. Denote the vector of prediction errors by $\widehat{U}_T(H) \equiv (\widehat{u}_T(1), \ldots, \widehat{u}_T(H))' \equiv \widehat{Y}_T(H) - Y_{T,H}$. Finally, $\widehat{\sigma}_T(h)$ denotes a prediction standard error, that is, a standard error for $\widehat{u}_T(h)$: it is an estimator of the unknown standard deviation of the random variable $\widehat{u}_T(h)$.

We further assume a generic method to compute a vector of bootstrap prediction errors $\widehat{U}^*_T(H) \equiv (\widehat{u}^*_T(1), \ldots, \widehat{u}^*_T(H))'$, based on artificial bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$ generated from an estimated probability mechanism $\widehat{\mathbb{P}}_T$. (The estimated probability mechanism has subscript T because it is a function of the observed data $\{y_1, \ldots, y_T\}$.) Such bootstrap forecast errors can be computed in many different ways. We shall not enter this debate here; the goal is to provide a generic procedure to construct a joint prediction region where application-specific details are up to the applied researcher. Finally, $\widehat{\sigma}^*_T(h)$ denotes a bootstrap prediction standard error, that is, a standard error for $\widehat{u}^*_T(h)$.

We now briefly illustrate these concepts. The observed data are $\{y_1, \ldots, y_T\}$. The applied researcher selects a suitable 'null' model, fits it to the data, and then uses the fitted model to make the predictions $\widehat{y}_T(h)$, for $h = 1, \ldots, H$. To be concrete, assume he uses an ARIMA model. The fitted model also provides prediction standard errors $\widehat{\sigma}_T(h)$. Next, the applied researchers generates bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$. To this end he can use a parametric bootstrap, based on the ARIMA model fitted from the original data; this would be a suitable approach if he believes that his null model is correctly specified. Alternatively, he can use a nonparametric time series bootstrap (say a blocks bootstrap or a sieve bootstrap); this would be a suitable approach if he believes that his null model might be misspecified.[4] Not making use of the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$, he computes forecasts $\widehat{y}^*_T(h)$ and prediction standard errors $\widehat{\sigma}^*_T(h)$. Finally, he computes $\widehat{u}^*_T(h) \equiv \widehat{y}^*_T(h) - y^*_{T+h}$.

Our high-level assumption below is based on the two vectors of standardized prediction errors $\widehat{S}_T(H) \equiv (\widehat{u}_T(1)/\widehat{\sigma}_T(1), \ldots, \widehat{u}_T(H)/\widehat{\sigma}_T(H))'$ and $\widehat{S}^*_T(H) \equiv (\widehat{u}^*_T(1)/\widehat{\sigma}^*_T(1), \ldots, \widehat{u}^*_T(H)/\widehat{\sigma}^*_T(H))'$, respectively. Denote the probability law under $\mathbb{P}$ of $\widehat{S}_T(H)|y_T, y_{T-1}, \ldots$ by $\widehat{J}_T$. Also denote the probability law under $\widehat{\mathbb{P}}_T$ of $\widehat{S}^*_T(H)|y^*_T, y^*_{T-1}, \ldots$ by $\widehat{J}^*_T$. In the asymptotic framework, T tends to infinity whereas H remains fixed.

Assumption 3.3.1. *$\widehat{J}_T$ converges in distribution to a non-random continuous limit law $\widehat{J}$. Furthermore, $\widehat{J}^*_T$ consistently estimates this limit law: $\rho(\widehat{J}_T, \widehat{J}^*_T) \to 0$ in probability for any metric ρ metrizing weak convergence.*

Expressed in words, Assumption 3.3.1 states that, as the sample size T increases, the conditional distribution of the vector of standardized bootstrap prediction errors $\widehat{S}^*_T(H)$ becomes a more and more reliable approximation to the (unknown) conditional distribution of the vector of true standardized prediction errors $\widehat{S}_T(H)$.

We next specify the forms of the joint prediction regions for $Y_{T,H}$, first for the two-sided case

[4] For an overview of nonparametric time series bootstrap methods, the reader is referred to Bühlmann (2002), Lahiri (2003), and Politis (2003).

and then for the one-sided case.

Some further notation is required. Suppose $X \equiv (x_1, \ldots, x_H)'$ is a vector with H components. First, for $k \in \{1, \ldots, H\}$, k-max(X) returns the k^{th} largest value of the x_h. So, if the elements x_h, $1 \leq h \leq H$, are ordered as $x_{(1)} \leq \cdots \leq x_{(H)}$, then k-max$(X) \equiv x_{(H-k+1)}$. Second, for $k \in \{1, \ldots, H\}$, k-min(X) returns the k^{th} smallest value of the x_h; that is, k-min$(X) \equiv x_{(k)}$. Third, $|X|$ denotes the vector $\big(|x_1|, \ldots, |x_H|\big)'$.

Let $d^{max}_{|\cdot|,1-\alpha}(k)$ denote the $1-\alpha$ quantile of the random variable k-max$\big(|\widehat{S}_T(H)|\big)$. Then a two-sided joint prediction region for $Y_{T,H}$ that exactly controls the k-FWE is given by

$$\big[\widehat{y}_T(1) \pm d^{max}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(1)\big] \times \cdots \times \big[\widehat{y}_T(H) \pm d^{max}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(H)\big] \ . \tag{3.12}$$

The implication is that the probability that the region (3.12) will contain at least $H-k+1$ elements of $Y_{T,H}$ is (at least) equal to $1-\alpha$ in finite samples. This property follows immediately from the definition of $d^{max}_{|\cdot|,1-\alpha}(k)$.

The problem is that this ideal region is not feasible, since the constant $d^{max}_{|\cdot|,1-\alpha}(k)$ is unknown. It has to be estimated in practice by $d^{max,*}_{|\cdot|,1-\alpha}(k)$, which is defined as the $1-\alpha$ quantile of the random variable k-max$\big(|\widehat{S}^*_T(H)|\big)$. This quantile can typically not be derived analytically, but it can be simulated to arbitrary precision from a sufficiently large number of bootstrap repetitions; see Algorithm 3.3.1 below.

Then a two-sided joint prediction region for $Y_{T,H}$ that asymptotically controls the k-FWE is given by

$$\big[\widehat{y}_T(1) \pm d^{max,*}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(1)\big] \times \cdots \times \big[\widehat{y}_T(H) \pm d^{max,*}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(H)\big] \ . \tag{3.13}$$

The implication is that the probability that the region (3.13) will contain at least $H-k+1$ elements of $Y_{T,H}$ is (at least) equal to $1-\alpha$ asymptotically.

The modifications to the one-sided case are as follows; we only present the feasible regions.

Let $d^{max,*}_{1-\alpha}(k)$ denote the $1-\alpha$ quantile of the random variable k-max$\big(\widehat{S}^*_T(H)\big)$. Then a one-sided lower joint prediction region for $Y_{T,H}$ that asymptotically controls the k-FWE is given by

$$\big[\widehat{y}_T(1) - d^{max,*}_{1-\alpha}(k) \cdot \widehat{\sigma}_T(1), \infty\big) \times \cdots \times \big[\widehat{y}_T(H) - d^{max,*}_{1-\alpha}(k) \cdot \widehat{\sigma}_T(H), \infty\big) \ . \tag{3.14}$$

Let $d^{min,*}_{\alpha}(k)$ denote the α quantile of the random variable k-min$\big(\widehat{S}^*_T(H)\big)$. Then a one-sided upper joint prediction region for $Y_{T,H}$ that asymptotically controls the k-FWE is given by

$$\big(-\infty, \widehat{y}_T(1) - d^{min,*}_{\alpha}(k) \cdot \widehat{\sigma}_T(1)\big] \times \cdots \times \big(-\infty, \widehat{y}_T(H) - d^{min,*}_{\alpha}(k) \cdot \widehat{\sigma}_T(H)\big] \ . \tag{3.15}$$

Note here that $d^{min,*}_{\alpha}(k)$ is generally a negative number so that, for each component h, the upper end of the corresponding interval is indeed larger than the prediction $\widehat{y}_T(h)$.

As is immediately clear from the definitions, the multipliers $d^{max,*}_{|\cdot|,1-\alpha}(k)$, $d^{max,*}_{1-\alpha}(k)$, and $d^{min,*}_{\alpha}(k)$ are each (weakly) monotonically decreasing in k. Consequently, the larger the value of k, the smaller in volume are the regions (3.13)–(3.15).

The following proposition formally establishes the asymptotic validity of these feasible bootstrap joint prediction regions.

Proposition 3.3.1. *Under Assumption 3.3.1 each of the joint prediction regions (JPRs)* (3.13)–(3.15) *for* $Y_{T,H}$ *satisfies*

$$\limsup_{T\to\infty} \; k\text{-}FWE \leq \alpha \;, \tag{3.16}$$

where

$$k\text{-}FWE \equiv \mathbb{P}\{\textit{At least } k \textit{ of the } y_{T+h} \textit{ not contained in the JPR}\} \;. \tag{3.17}$$

The following algorithm details how to compute the three multipliers $d^{max,*}_{|\cdot|,1-\alpha}(k)$, $d^{max,*}_{1-\alpha}(k)$, and $d^{min,*}_{\alpha}(k)$ in practice. The algorithm assumes a generic bootstrap method, chosen by the applied researcher, to generate bootstrap data and standardized bootstrap prediction errors. In particular, such a bootstrap method is based on an estimated probability mechanism $\widehat{\mathbb{P}}_T$.

Algorithm 3.3.1 (Computation of the JPR Multipliers; Univariate Case)**.**

1. Generate bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$ from $\widehat{\mathbb{P}}_T$.
2. Not making use of the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$, compute forecasts $\widehat{y}^*_T(h)$ and prediction standard errors $\widehat{\sigma}^*_T(h)$.
3. Compute bootstrap prediction errors $\widehat{u}^*_T(h) \equiv \widehat{y}^*_T(h) - y^*_{T+h}$.
4. Compute standardized bootstrap prediction errors $\widehat{s}^*_T(h) \equiv \widehat{u}^*_T(h)/\widehat{\sigma}^*_T(h)$ and let $\widehat{S}^*_T(H) \equiv \big(\widehat{s}^*_T(1), \ldots, \widehat{s}^*_T(H)\big)'$.
5. Compute $k\text{-max}^*_{|\cdot|} \equiv k\text{-max}\big(\big|\widehat{S}^*_T(H)\big|\big)$, $k\text{-max}^* \equiv k\text{-max}\big(\widehat{S}^*_T(H)\big)$, and $k\text{-min}^* \equiv k\text{-min}\big(\widehat{S}^*_T(H)\big)$.
6. Repeat this process B times, resulting in statistics $\{k\text{-max}^*_{|\cdot|,1}, \ldots, k\text{-max}^*_{|\cdot|,B}\}$, $\{k\text{-max}^*_1, \ldots, k\text{-max}^*_B\}$, and $\{k\text{-min}^*_1, \ldots, k\text{-min}^*_B\}$.
7. Compute the corresponding empirical quantiles:
   - 7.1 $d^{max,*}_{|\cdot|,1-\alpha}(k)$ is the empirical $1-\alpha$ quantile of the statistics $\{k\text{-max}^*_{|\cdot|,1}, \ldots, k\text{-max}^*_{|\cdot|,B}\}$.
   - 7.2 $d^{max,*}_{1-\alpha}(k)$ is the empirical $1-\alpha$ quantile of the statistics $\{k\text{-max}^*_1, \ldots, k\text{-max}^*_B\}$.
   - 7.3 $d^{min,*}_{\alpha}(k)$ is the empirical α quantile of the statistics $\{k\text{-min}^*_1, \ldots, k\text{-min}^*_B\}$.

In an application, the number of bootstrap repetitions, B, should be chosen as large as possible; at the very least $B \geq 1,000$.

Remark 3.3.1. Proposition 3.3.1 only addresses asymptotic consistency. It does not address finite-sample performance. To ensure best-possible finite-sample performance the applied researcher should make an effort to match the bootstrap distribution $\widehat{J}^*_T$ as close as possible to the true distribution $\widehat{J}_T$. How this is to be done in detail depends on the particular bootstrap method chosen by the applied researcher. Many papers have been written on this problem already; for example, see De Gooijer and Hyndman (2006, Section 12).

We confine ourselves to the general statement that model parameters which have to be estimated from the original data $\{y_1, \ldots, y_T\}$ to compute the forecasts $\widehat{y}_T(h)$ and the prediction

standard errors $\widehat{\sigma}_T(h)$ should be re-estimated from the bootstrap data $\{y_1^*, \ldots, y_T^*\}$ to compute the forecasts $\widehat{y}_T^*(h)$ and the prediction standard errors $\widehat{\sigma}_T^*(h)$. It may be tempting, say in order to save computing time, to simply use the estimated model parameters from the original data $\{y_1, \ldots, y_T\}$ to compute the forecasts $\widehat{y}_T^*(h)$ and the prediction standard errors $\widehat{\sigma}_T^*(h)$. But such an approach does not reflect the fact that the true model parameters are unknown and generally leads to bootstrap prediction errors that are too small in magnitude. ■

3.3.2 Multivariate Time Series

Compared to the special case of a univariate time series, the methodology does not change in any fundamental way in the general case of a multivariate time series, as in the case of VAR forecasting. Mainly, the notation becomes more complex.

One observes a K-variate time series $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$ generated from a true probability mechanism $\mathbb{P}$ and wishes to predict the next stretch of H observations for a particular component of $\boldsymbol{z}_t$. Assume without loss of generality that one wishes to predict the first component of $\boldsymbol{z}_t$ and write $\boldsymbol{z}_t \equiv (y_t, x_{2,t}, \ldots, x_{K,t})'$.

In this more general case, the forecast of y_{T+h}, denoted by $\widehat{y}_T(h)$ again, will be function of $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$ instead of a function of $\{y_1, \ldots, y_T\}$ only; and similarly for the prediction standard errors $\widehat{\sigma}_T(h)$.

Artificial bootstrap data $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_T^*, \boldsymbol{z}_{T+1}^*, \ldots, \boldsymbol{z}_{T+H}^*\}$ are generated from an estimated probability mechanism $\widehat{\mathbb{P}}_T$. In particular, K-variate VAR models appear a popular choice to this end with applied researchers; more generally, SVAR, VECM, or SVECM models can also be used; for example, see Lütkepohl (2005).

Denote $\boldsymbol{z}_t^* \equiv (y_t^*, x_{2,t}^*, \ldots, x_{H,t}^*)'$. The forecast of y_{T+h}^*, denoted by $\widehat{y}_T^*(h)$ again, will be a function of $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_T^*\}$ instead of a function of $\{y_1^*, \ldots, y_T^*\}$ only; and similarly for the prediction standard errors $\widehat{\sigma}_T^*(h)$.

Assumption 3.3.1 continues to be based on the two vectors of standardized prediction errors $\widehat{S}_T(H) \equiv (\widehat{u}_T(1)/\widehat{\sigma}_T(1), \ldots, \widehat{u}_T(H)/\widehat{\sigma}_T(H))'$ and $\widehat{S}_T^*(H) \equiv (\widehat{u}_T^*(1)/\widehat{\sigma}_T^*(1), \ldots, \widehat{u}_T^*(H)/\widehat{\sigma}_T^*(H))'$, respectively. Only that now, more generally, $\widehat{J}_T$ denotes the probability law under $\mathbb{P}$ of $\widehat{S}_T(H)|\boldsymbol{z}_T, \boldsymbol{z}_{T-1}, \ldots$; and $\widehat{J}_T^*$ denotes the probability law under $\widehat{\mathbb{P}}_T$ of $\widehat{S}_T^*(H)|\boldsymbol{z}_T^*, \boldsymbol{z}_{T-1}^*, \ldots$

Having detailed how the quantities of interest are defined and computed in the more general case, the methodology outlined in the case of a univariate time series applies verbatim. The various forms of the joint prediction regions are still given by (3.13)–(3.15) and Proposition 3.3.1 continues to hold.

The following algorithm details how to compute the three multipliers $d_{|\cdot|,1-\alpha}^{max,*}(k)$, $d_{1-\alpha}^{max,*}(k)$, and $d_{\alpha}^{min,*}(k)$ in practice. The algorithm assumes a generic bootstrap method, chosen by the applied researcher, to generate bootstrap data and standardized bootstrap prediction errors. In particular, such a bootstrap method is based on an estimated probability mechanism $\widehat{\mathbb{P}}_T$.

Algorithm 3.3.2 (Computation of the JPR Constants; Multivariate Case)**.**

1. Generate bootstrap data $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_T^*, \boldsymbol{z}_{T+1}^*, \ldots, \boldsymbol{z}_{T+H}^*\}$ from $\widehat{\mathbb{P}}_T$.
2. Not making use of the stretch $\{\boldsymbol{z}_{H+1}^*, \ldots, \boldsymbol{z}_{T+H}^*\}$, compute forecasts $\widehat{y}_T^*(h)$ and prediction standard errors $\widehat{\sigma}_T^*(h)$.
3. Identical to Algorithm 3.3.1.

⋮

7. Identical to Algorithm 3.3.1.

3.3.3 Comparison with Previous Methods

Jordà and Marcellino (2010) propose an alternative 'asymptotic' method to construct a joint prediction region for $Y_{T,H}$ that controls the FWE.[5] It is based on the assumption that

$$\sqrt{T}\big(\widehat{Y}_T(H) - Y_{T,H}|\boldsymbol{z}_T, \boldsymbol{z}_{T-1}, \ldots\big) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Xi}_H) \ , \tag{3.18}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and on the availability of a consistent estimator $\widehat{\boldsymbol{\Xi}}_H \xrightarrow{\mathbb{P}} \boldsymbol{\Xi}_H$, where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability.

The proposed joint prediction region is given by

$$\widehat{Y}_T(H) \pm P \left[\sqrt{\frac{\chi^2_{h,1-\alpha}}{h}} \right]_{h=1}^{H} \ , \tag{3.19}$$

where P is the lower-triangular Cholesky decomposition of $\widehat{\boldsymbol{\Xi}}_H/T$, satisfying $PP' = \widehat{\boldsymbol{\Xi}}_H/T$, and the quantity to the right of P is a $H \times 1$ vector whose h^{th} entry is given by $\sqrt{\chi^2_{h,1-\alpha}/h}$. This approach is problematic for several reasons.

First, assumption (3.18) implies that the conditional distribution of the vector of prediction errors $\widehat{U}_T(H) \equiv \widehat{Y}_T(H) - Y_{T,H}$ is approximately multivariate normal with mean zero, at least for large T. This appears overly strict. The conditional distribution of a prediction error depends on the conditional distribution of the random variable to predicted. If the latter distribution is non-normal, which is the case in many applications, then the former distribution is generally non-normal as well.

Second, assumption (3.18) implies in addition that the conditional covariance matrix of the vector of prediction errors $\widehat{U}_T(H) \equiv \widehat{Y}_T(H) - Y_{T,H}$ vanishes asymptotically. This appears unrealistic. While, under mild regularity conditions, the variance of an estimator of a population parameter vanishes asymptotically, the same is not true for the variance of a prediction error. Even if all model parameters are known, a future observation cannot be predicted perfectly because of its random nature.

[5] They use the term *joint confidence region* instead of *joint prediction region*.

Remark 3.3.2. To illustrate the first two points, consider the simple AR(1) model

$$y_t = \nu + \rho y_{t-1} + \epsilon_t \ , \tag{3.20}$$

where $|\rho| < 1$ and the errors $\{\epsilon_t\}$ are independent and identically distributed (i.i.d.) with mean zero and finite variance σ_ϵ^2. At time T, the forecast of y_{T+1} is given by

$$\widehat{y}_T(1) \equiv \widehat{\nu} + \widehat{\rho} y_T \ , \tag{3.21}$$

where $\widehat{\nu}$ and $\widehat{\rho}$ are suitable, consistent estimators of ν and ρ. The forecast error is given by

$$\widehat{u}_T(1) = \widehat{\nu} + \widehat{\rho} y_T - y_{T+1} \ . \tag{3.22}$$

As T tends to infinity, the conditional distribution of $\widehat{u}_T(1)$ converges weakly to the unconditional distribution of $-\epsilon_{T+1}$ (which does not depend on T). This distribution is neither necessarily normal nor does its variance vanish. As a result, assumption (3.18) does not hold in this simple example. ■

Third, Jordà and Marcellino (2010) initially consider the following rectangular joint prediction region:

$$\widehat{Y}_T(H) \pm P \left[\sqrt{\frac{\chi^2_{H,1-\alpha}}{H}} \, \mathbf{1}_H \right] \ , \tag{3.23}$$

where $\mathbf{1}_H$ is a $H \times 1$ vector of ones. It is derived by an application of Bowden's (1970) lemma to an elliptical joint prediction region based on Scheffé's method:

$$\left\{ \widetilde{Y} : T(\widehat{Y}_T(H) - \widetilde{Y})' \widehat{\mathbf{\Xi}}_H^{-1} (\widehat{Y}_T(H) - \widetilde{Y}) \leq \chi^2_{H,1-\alpha} \right\} \ . \tag{3.24}$$

As we have explained above, deriving a rectangular joint confidence region from an initial joint confidence region of elliptical form is suboptimal in terms of the volume of the rectangular joint confidence region.

Fourth, Jordà and Marcellino (2010) arrive at their final joint prediction region (3.19) by 'refining' the initial joint prediction region (3.23) using a step-down recursive procedure that is entirely ad-hoc and lacks a theoretical justification.

Since there is no proof of asymptotic validity, under realistic conditions, of the method proposed by Jordà and Marcellino (2010), the method is not trustworthy to use in practice.

Staszewska-Bystrova (2010) proposes an alternative bootstrap method to construct a joint prediction region for $Y_{T,H}$ that controls the FWE. In a nutshell, the method works as follows. Conditional on the observed data, one generates B bootstrap path-forecasts $\widehat{Y}_T^{*,b}(H)$, for $b = 1, \ldots, B$. One then discards αB of these bootstrap path-forecasts: namely those $\widehat{Y}_T^{*,b}(H)$ that are 'furthest' away from the original path-forecast $\widehat{Y}_T(H)$, where the distance between two $H \times 1$ vectors is measured by the Euclidian distance.[6] Finally, the joint prediction region

[6] Staszewska-Bystrova (2010) also considers other distance measures, but concludes that the Euclidean distance seems to work best.

is given by the envelope of the remaining $(1-\alpha)B$ bootstrap path-forecasts. Although this methods seems to perform well in some simulation studies, three criticism apply.

First, the method is purely heuristic. No proof of asymptotic validity, under some suitable high-level assumption, is provided.

Second, the method seems restricted to (V)AR models, since it uses the backward representation of a (V)AR model to generate the bootstrap path-forecasts; see Thombs and Schucany (1990) for an early use of this representation in AR models. As an additional restriction, a problem of the backward representation when the forward errors are non-normal, is that even if the forward errors are independent, the backward errors are not independent, but merely uncorrelated; Pascual et al. (2001) point this out already. Hence, using Efron's bootstrap on the residuals in the backward representation, as proposed by Staszewska-Bystrova (2010), may not be generally valid.

Third, the method is in the spirit of Efron's percentile method that amounts to "looking up the wrong tails of a distribution"; see Hall (1992, Section 1.3 and 3.4) for a discussion. Theoretical arguments suggest that such a method can only work well when the conditional distribution of the vector of forecast errors is symmetric around zero, as would be the case for a multivariate normal distribution. The performance of the method may suffer when prediction errors are, conditionally, skewed or have non-zero mean. Staszewska-Bystrova (2010) only considers normal error terms with mean zero in the data generating processes (DGPs) of her simulation study. On the other hand, the joint prediction regions we propose in Subsections 3.3.1 and 3.3.2 are based on Hall's percentile-t method, which has a sound theoretical foundation and is more generally valid than Efron's percentile method; again see Hall (1992, Sections 1.3 and 3.4).

Last but not least, it is not clear whether the methods of Jordà and Marcellino (2010) or Staszewska-Bystrova (2010) can be generalized to construct a joint prediction region for $Y_{T,H}$ that controls the k-FWE for $k \geq 2$. By offering a method to construct rectangular joint prediction regions for $Y_{T,H}$ that control the k-FWE for arbitrary $k \geq 1$, we provide applied researchers with a more flexible and versatile tool.

3.4 Monte Carlo Study

We analyze multiple error type I properties of our joint prediction regions and competing forecast bands across a range of relevant data generating processes (DGPs). Bear in mind that technically, the methods competing with our k-FWE JPRs are not joint prediction regions; calling them forecast bands as a somewhat flexible term seems appropriate.

The multiple error type I clearly is the realized k-FWE. We report k-FWE coverage; thus the closer the realized coverage $1-k$-FWE is to the desired level $1-\alpha_{k\text{-FWE}}$, the better is the associated forecast band. There are no competing methods to our k-FWE JPR bands for

$k > 1$.

3.4.1 Simulation Setup

Undoubtedly, how well a chosen model approximates the unobservable data generating process is an important question for each empirical application. If a badly fitting model is chosen in a particular application, the properties of the resulting forecast bands are arbitrary with respect to the data generating process; using a nonparametric bootstrap instead of a model-based one may be a means to mitigate this problem. In what follows, we focus exclusively on how well the different forecast band methods perform with respect to a fixed approximation quality of the chosen model. To this end, we use different AR(p) data generating processes and models thereof with fixed or BIC selected lag length.

Before describing which data generating processes we look at in section 3.4.2, let us explain how the simulation for one specific AR(p) data generating process is carried out. We make use of AR(p) data generating processes

$$y_t = \rho_1 \cdot y_{t-1} + \ldots + \rho_p \cdot y_{t-p} + \epsilon_t, \quad \boldsymbol{\rho} = [\rho_1, \ldots, \rho_p] \tag{3.25}$$

as follows. We first simulate $M = 1000$ time series of length $T + H$ from this DGP, denoted as $y_{T+H}^{(m)}$ for each Monte Carlo run $m = 1, \ldots, M$.

Second, given the observations simulated from the DGP up to T, denoted as $y_{1:T}^{(m)}$, we fit an AR(p) model including a constant, yielding an estimate $\hat{\boldsymbol{\rho}}^{(m)}$ of the true $\boldsymbol{\rho}$. This $\hat{\boldsymbol{\rho}}^{(m)}$ is bias-corrected, resulting in $\tilde{\boldsymbol{\rho}}^{(m)}$.

Third, using this estimate $\tilde{\boldsymbol{\rho}}^{(m)}$, the predictions from $h = 1$ to H in vector $\widehat{y}_H^{(m)}$ as well as the associated vector of H standard errors are estimated on each run m. To compute the Scheffe bands, one needs to compute the covariance matrix of $\widehat{y}_H^{(m)}$.

Fourth, using $B = 1000$ bootstrap runs on each data set $y_{1:T}^{(m)}$, four different types of forecast bands $FCB(\widehat{y}_H^{(m)})$ are computed

1. Naive bands by joining the marginal bootstrap prediction intervals over $h = 1, \ldots, H$
2. Our k-FWE joint prediction regions (k-FWE JPRs)
3. Scheffe forecast bands as in Jordà and Marcellino (2010)
4. heuristic neighboring paths (NP) forecast bands from Staszewska-Bystrova (2010)

We use the studentized Pascual et al. (2001) bootstrap scheme, which essentially conditions every bootstrap prediction on the last p observed data points $y_{(T-p+1):T}^{(m)}$. To get an impression what these forecast bands look like on one particular Monte Carlo data set m, see Figure 4.2 and 4.3.

Fifth, on each Monte Carlo run m, we compare the forecast bands to future realizations of the process, which we call continuations in accordance with Clements and Taylor (2001). To increase the accuracy of our results, we follow them in comparing $FCB(\widehat{y}_H^{(m)})$ not only to one continuation $y_{(T+1):(T+H)}^{(m)}$; we compare $FCB(\widehat{y}_H^{(m)})$ to $c = 1, \ldots, 100$ continuations $y_{(T+1):(T+H)}^{(m),(c)}$ that are all based on the last p observations $y_{(T-p+1):(T)}^{(m)}$ of Monte Carlo data set m. If the simulated continuation $y_{(T+1):(T+H)}^{(m),(c)}$ lies outside the forecast band $FCB(\widehat{y}_H^{(m)})$ at k or more prediction horizons h, then a k familywise error k-FWE was committed on this particular continuation c of Monte Carlo data set m. Thus, the mean realized k-FWE$^{(m)}$ on Monte Carlo data set m is

$$\text{k-FWE}^{(m)} = \frac{1}{100}\sum_{c=1}^{100} \mathbf{1}\left[\sum_{h=1}^{H} \mathbf{1}[y_{(T+1):(T+H)}^{(m),(c)} \notin FCB(\widehat{y}_H^{(m)})] \geq k\right] \tag{3.26}$$

The realized coverage $1-k$-FWE is one minus the arithmetic mean over all $M = 1000$ Monte Carlo simulation runs

$$1 - k\text{-FWE} = 1 - \frac{1}{M}\sum_{m=1}^{M} k\text{-FWE}^{(m)} \tag{3.27}$$

The Scheffe bands as such do not require bootstrapping, the modus operandi propagated in Jordà and Marcellino (2010) is to rely on asymptotic normality[7]. To compute these Scheffe bands, we include parameter estimation uncertainty in the covariance matrix of the predictions, as described in the appendix of Jordà and Marcellino (2010)'s paper. The Staszewska-Bystrova (2010) NP heuristic makes use of the same bootstrap scheme that we use for computing our k-FWE JPR bands, except that no studentization is used for the NP bands.

The average volume of a forecast band over the $M = 1000$ Monte Carlo runs is

$$\text{Volume} = \frac{1}{M}\sum_{m=1}^{M}\prod_{h=1}^{H}\left|\text{Upper}_h(FCB(\widehat{y}_H^{(m)})) - \text{Lower}_h(FCB(\widehat{y}_H^{(m)}))\right|, \tag{3.28}$$

where $\text{Upper}_h(FCB(.))$ and $\text{Lower}_h(FCB(.))$ denotes the upper and lower ends of the forecast bands at horizon h, respectively. For the sake of brevity, we do not report average volumes within the results tables of Section 3.4.3; they are available from the authors on request. Nonetheless, there is the short Subsection 3.4.3 on volumes as well as the two Figures 4.2 and 4.3 that exemplify some volume properties.

3.4.2 Data Generating Processes in Simulations

Fixed Lag Length

For the fixed lag length, we use AR(1) and AR(2) processes. The error term ϵ_t is distributed as follows: $\epsilon_t \sim N(0,1)$, or $\epsilon_t \sim \frac{t_3}{\sqrt{3}}$, or $\epsilon_t \sim \frac{\chi_3^2-3}{\sqrt{2\cdot 3}}$. Note that all three errors have mean zero

[7]The Scheffe results in our simulation study do not change considerably by using their finite-sample Scheffe bands, comprising the F-distribution for computing the Scheffe critical values instead of the χ^2-distribution

and variance one to facilitate comparisons of results across different error distributions. The data size T is either 100 or 400. We choose forecast horizons $H \in \{6, 12, 24\}$.

The bias correction of $\hat{\boldsymbol{\rho}}$ is as in Kilian (1998) using the computational shortcut. Hence, only $B + B$ bootstrap runs are necessary for this bias correction, not $B \cdot B$ as in his original bootstrap-after-bootstrap bias correction.

For AR(1), the autoregressive coefficient $\boldsymbol{\rho}$ is in $\{0.9, 0.5, -0.5, -0.9\}$. The AR(1) results for data sets of length $T = 100$ are reported in Table 3.1, the AR(1) results for $T = 400$ are in Table 3.2.

For AR(2), $\boldsymbol{\rho}$ is one of the following $\{[1.75, -0.85], [1.25, -0.75], [-0.65, 0.15], [-0.7, -0.2]\}$. The AR(2) results for $T = 100$ are in Table 3.3, the AR(2) results for $T = 400$ are in Table 3.4.

BIC Selected Lag Length

For the BIC selected lag length, we focus on the AR(2) process with $\epsilon_t \sim N(0, 1)$. Again, the data size T is either 100 or 400 with forecast horizons $H \in \{6, 12, 24\}$. As before, $\boldsymbol{\rho}$ is one of the following $\{[1.75, -0.85], [1.25, -0.75], [-0.65, 0.15], [-0.7, -0.2]\}$. These results are summarized in Table 3.5.

We bias-correct $\hat{\boldsymbol{\rho}}$ according to the closed-form bias in White (1961) that facilitates individual BIC lag length selection on each bootstrap data set b. In contrast, the $(B + B)$-sized shortcut version of Kilian (1998)'s bias correction makes automatic lag selection cumbersome. His original computing-intensive $(B \cdot B)$-sized bias correction allows for automatic lag selection on each bootstrap data set, which seems too computationally expensive, however. Maybe this is one of the reasons why the Clements and Taylor (2001) study on bootstrapping predictions is entirely based on fixed lag lengths.

For the sake of brevity, the results using the AIC and AICc criterion for lag length selection are not reported; the comparative pattern between the methods remains the same.

3.4.3 Results

Fixed Lag Length

We discuss the AR(1) $T = 100$ results in Table 3.1 at the same time as the AR(1) $T = 400$ results in Table 3.2. In each of these tables, the realized coverage $1-k$-FWE is reported, which should be as close as possible to the desired coverage $1 - \alpha_{k\text{-FWE}} = 0.9$. Corresponding to a DGP, each of the $\boldsymbol{\rho} \in \{0.9, 0.5, -0.5, -0.9\}$ is placed in one of the four blocks of rows separated by a horizontal line. For each DGP, we report the naive joining of marginal prediction intervals, our k-FWE JPR bands for $k = 1, 2, 3$, the Scheffe bands, and the NP heuristic bands. Each of the error terms $\epsilon_t \sim N(0, 1)$, $\epsilon_t \sim \frac{t_3}{\sqrt{3}}$, and $\epsilon_t \sim \frac{\chi^2_3 - 3}{\sqrt{2 \cdot 3}}$ corresponds to a block of three columns

labeled accordingly. Within each of these blocks of three columns, we report the realized k-FWE coverage for forecast horizons $H \in \{6, 12, 24\}$.

First, note that joining H marginal prediction intervals at marginal level $\alpha = 0.1$ is unrelated to constructing a joint prediction region at multiple level $\alpha_{k\text{-FWE}} = 0.1$ for $k = 1$, which is denoted as $\alpha_{1\text{-FWE}}$ in the following. Such a naive approach leads to seriously undercovering joint prediction regions.

Second, the realized coverage of the 1-FWE JPR bands is closest to the nominal level in all cases. The more liberal 2-FWE and 3-FWE JPR bands achieve realized coverage close to the desired level. There are no competitors for these two more liberal forecast bands.

Third, the Scheffe band's 1-FWE coverage is quite close to the nominal level for $\boldsymbol{\rho} = 0.9$. For $\boldsymbol{\rho} \in [0.5, -0.5, -0.9]$, however, the Scheffe 1-FWE coverage monotonically decreases to zero. In its most extreme form, the realized 1-FWE coverage of the Scheffe bands for $\boldsymbol{\rho} = -0.9$ is close to zero. This deficient coverage is attributable to the the critical value of the Scheffe forecast bands becoming smaller for increasing H instead of becoming larger, violating a basic insight from multiple testing that applies to joint prediction regions as well. It can be proved that the 1-FWE size distortion of Scheffe bands in the AR(1) case is monotonically decreasing the larger $\boldsymbol{\rho} < 1$ gets. The proof is available from the authors on request (see Essay 4). Jordà and Marcellino (2010) partly report this deficient property for AR(1) processes; for example, see the bottom of their Table III for $H = 12$ as $\boldsymbol{\rho} = 0.5$ approaches $\boldsymbol{\rho} = 0.9$.

Fourth, the NP heuristic bands achieve realized 1-FWE coverages quite close to the nominal level, except for the $T = 100$ $\epsilon_t \sim t_3$ $H = 24$ case.

Not surprisingly, the coverages of the k-FWE forecast bands based on $T = 400$ data sets are uniformly better than the ones for $T = 100$. The same holds for the Scheffe bands. We observe the same pattern for the NP heuristic bands, although it is not theoretically clear why this holds.

The AR(2) results in Table 3.3 and Table 3.4 are organized as in the AR(1) case. Again, the 1-FWE JPR bands achieve realized coverage closest to the desired level.

The more liberal 2-FWE and 3-FWE JPR bands achieve good k-FWE coverage; there are no competing methods.

The Scheffe band's coverage is worse than in the AR(1) case, especially for the $\boldsymbol{\rho} = [1.75, -0.85]$ DGP, on which the Scheffe bands work well in the AR(1) case. We observe the same deficient 1-FWE coverage properties of Scheffe bands as $\boldsymbol{\rho}$ gets more negative[8]. Again, the NP heuristic's 1-FWE coverage is too low for the $T = 100$ $\epsilon_t \sim t_3$ $H = 24$ case. As before, the $T = 400$ forecast bands fare better than the ones based on $T = 100$ data sets.

[8] In the sense that $\rho_1 + \rho_2$ decreases

BIC Selected Lag Length

Table 3.5 summarizes results for BIC selected lag lengths. There are only two blocks of three columns, because we merely look at standard normal errors. The first block of three columns contains the $T = 100$ results, the second block of three columns contains the $T = 400$ results.

Essentially, the comparative pattern observed in the fixed lag length AR(1) and AR(2) setups remain unchanged. That is, the 1-FWE JPR bands achieve coverage closest to the nominal level. The 2-FWE and 3-FWE JPR bands yield good k-FWE coverage. The Scheffe bands seriously undercover, except for some cases in the $\boldsymbol{\rho} = [1.75, -0.85]$ DGP. The deficient Scheffe band coverage as $\boldsymbol{\rho}$ gets more negative is as before. The NP heuristic bands achieve realized 1-FWE coverage close to the desired one; in the $\boldsymbol{\rho} = [1.75, -0.85]$ $H = 24$ case they are somewhat conservative. The forecast bands based on $T = 400$ data sets achieve better coverage than those based on $T = 100$ data sets.

The comparative pattern between the considered methods remains unchanged for AIC and AICc selected lag lengths. The realized coverages change by one percentage point at most; these results are not reported for the sake of brevity.

Volumes of the Forecast Bands

The average volume of the forecast error bands indicates multiple error type II properties. Conceptually, if the realized coverages $1-k$-FWE of two methods are comparably close to the nominal level $1 - \alpha_{k\text{-FWE}}$, the forecast band with the lower average volume is superior[9]. For the considered methods, it is redundant to report multiple error type II properties: the closer the 1-FWE coverage of a method is to the nominal level, the larger is the average volume of the associated forecast band. The notable exception to this rule is the Scheffe method. Figure 4.2 and Figure 4.3, which are part of Essay 4, display forecast bands $FCB(\widehat{y}_H^{(m)})$ for the AR(1) setup for positive and negative $\boldsymbol{\rho}$, respectively, together with the first continuation $y_{(T+1):(T+H)}^{(m),(c=1)}$ on one particular Monte Carlo data set ($m = 6$). The volumes observed on this Monte Carlo data set are representative of the average volume pattern observed on all DGPs over all simulation runs.

That is, the volume of our 1-FWE JPR bands and the NP heuristic bands are comparably small. As one gets more liberal in increasing the k within the k-FWE JPR band, its volume becomes smaller. For $\boldsymbol{\rho} = 0.9$, the volume of the Scheffe bands is too large for the coverage it achieves. For negative $\boldsymbol{\rho}$, the volume of the Scheffe bands is far too small, leading to the reported deficient 1-FWE coverage properties.

Similar plots for the fixed lag length AR(2) and BIC AR(2) exemplify the same pattern.

[9] At the present time, there are only competitors to our 1-FWE JPR bands

3.5 Empirical Application

To come.

3.6 Conclusions

Many economic and financial applications require the forecast of a random variable of interest over several periods into the future, that is, one needs to forecast an entire future path. In addition to the resulting path-forecast, one often would also like to compute a corresponding joint prediction region. Such a region is supposed to contain the entire future path with a prespecified probability $1 - \alpha$.

In this paper, we have proposed bootstrap joint prediction regions of three different shapes: one-sided lower, one-sided upper, and two-sided. In this way, the applied researcher can choose the most suitable shape for the task at hand. Furthermore, the joint prediction regions are completely generic in that they allow the applied researcher to select whichever methods are deemed most appropriate by him to make forecasts, compute prediction standard errors, and generate bootstrap data.

Compared to two previous proposals in the literature, our bootstrap joint prediction regions have two important advantages. First, they are proven to be asymptotically consistent under a realistic, mild high-level assumption. Second, they enjoy superior finite-sample properties, as demonstrated via Monte Carlo simulations.

As an additional bonus, we also offer generalized joint prediction regions obtained by the bootstrap. Such regions are not required to contain the entire future path (with prespecified probability $1 - \alpha$) but only the entire future path up to a small, user-defined number of elements (with prespecified probability $1 - \alpha$). If the maximum forecast horizon is large, it may be deemed acceptable by the applied researchers that a small number, like one or two, of elements of the future path fall outside the joint prediction region with a prespecified small probability α. In return, he will then obtain a smaller and more informative region.

3.A Proofs

PROOF OF PROPOSITION 3.3.1: We prove the stated result for the joint prediction region (3.14). The proofs for the joint prediction regions (3.13) and (3.15) are completely analogous.

Let $\widehat{L}_T$ denote a random variable with distribution $\widehat{J}_T$ and let $\widehat{L}$ denote a random variable with distribution $\widehat{J}$. By Assumption 3.3.1 and the continuous mapping theorem, k-max$(\widehat{L}_T)$ converges weakly to k-max$(\widehat{L})$, whose distribution is continuous. Our notation implies that the conditional sampling distribution under $\mathbb{P}$ of k-max$(\widehat{S}_T(H))$ is identical to the distribution of k-max$(\widehat{L}_T)$. By similar reasoning, the conditional sampling distribution under $\widehat{\mathbb{P}}_T$ of k-max$(\widehat{S}^*_T(H))$ also converges weakly to the distribution of k-max$(\widehat{L})$. To then show that

$$\mathbb{P}\big\{k\text{-max}(\widehat{S}_T(H)) \leq d^*_{1-\alpha}(k)\big\} \to 1-\alpha \tag{3.29}$$

is similar to the proof of Theorem 1 of Beran (1984).

Since by definition of the k-FWE and the construction of the joint prediction region (3.14),

$$k\text{-FWE} = 1 - \mathbb{P}\big\{k\text{-max}(\widehat{S}_T(H)) \leq d^*_{1-\alpha}(k)\big\} \ , \tag{3.30}$$

the proof that the stated result (3.16) holds for the joint prediction region (3.14) now follows immediately from (3.29). ■

3.B Generalized Error Rates, Multiple Testing, and Joint Confidence/Prediction Regions

The goal of this appendix is to explain why control of the *false discovery rate* (FDR) is actually equivalent to control of the *familywise error rate* (FWE) in the context of joint confidence regions and joint prediction regions.

In doing so, we first need to discuss some concepts from the literature on *multiple testing*. In a multiple testing problem one considers H individual hypotheses of the kind

$$H_{0,h} : \mu_h = \mu_{0,h} \quad \text{vs.} \quad H_{1,h} : \mu_h \neq \mu_{0,h} \ . \tag{3.31}$$

(For concreteness, we consider two-sided hypotheses here; one could also consider one-sided hypotheses instead.) The goal is to make individual decisions, in terms of rejecting or not, concerning each $H_{0,h}$ while controlling a prespecified error rate.

Denote by $I(\mathbb{P})$ the set of true null hypotheses, that is,

$$I(\mathbb{P}) \equiv \{h : H_{0,h} \text{ is true}\} \ . \tag{3.32}$$

The most stringent error rate is the *familywise error rate* (FWE), defined as the probability of rejecting at least one true null hypothesis:

$$\text{FWE} \ \equiv \mathbb{P}\{\text{Reject at least one of the } H_{0,h} : h \in I(\mathbb{P})\} \ . \tag{3.33}$$

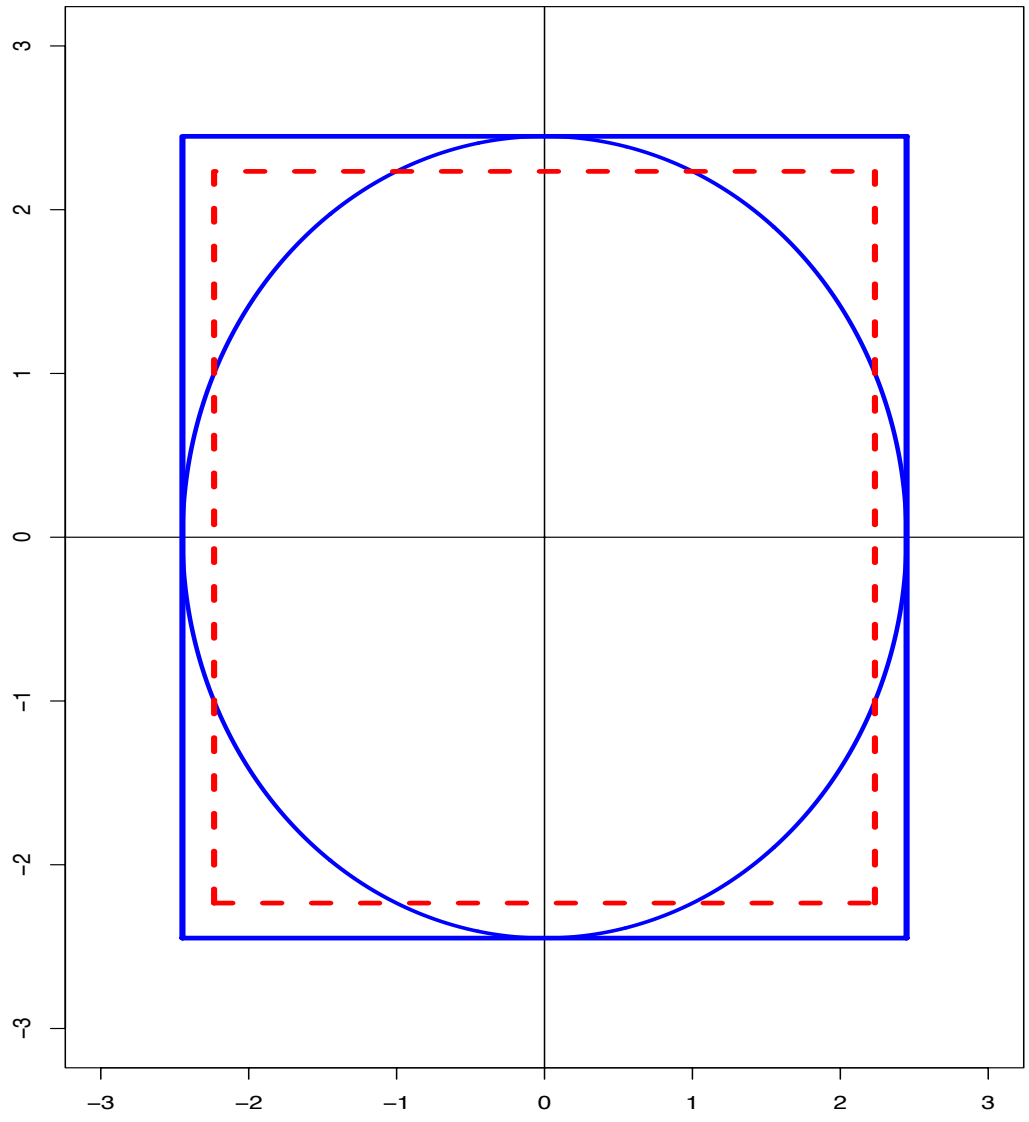


Figure 3.1: An illustration of Remark 3.2.1. One observes $\widehat{\boldsymbol{\mu}} = Y = (0.0, 0.0)$ and wishes to construct a joint confidence region for $\boldsymbol{\mu}$ with confidence level $1 - \alpha = 0.95$. The solid ellipse is the Scheffé joint confidence region: a circle with radius 2.45. The solid rectangle is the implied (that is, projected on the axes) rectangular joint confidence region: a square with half length 2.45. The dashed rectangle is the 'direct' rectangular joint confidence region: a square with half length 2.24.

Nominal Coverage $1 - \alpha_{k\text{-FWE}} = 90\%$

Empirical $1-k$-FWE	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2_3$		
$\boldsymbol{\rho} = 0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	66.7	53.3	38.2	71.1	58.6	43.6	69.3	56.2	41.3
k-FWE JPR ($k=1$)	90.3	90.0	89.7	90.1	89.5	88.8	89.8	90.3	90.1
k-FWE JPR ($k=2$)	90.2	89.6	88.9	90.6	89.3	88.2	90.1	90.0	89.9
k-FWE JPR ($k=3$)	89.8	89.3	89.2	89.9	89.7	88.3	90.3	89.3	90.2
Scheffé ($k=1$)	86.6	85.9	84.9	86.9	85.5	83.6	88.7	87.7	86.5
NP Heuristic ($k=1$)	90.1	90.6	90.9	88.9	88.1	86.6	90.6	90.7	90.4
$\boldsymbol{\rho} = 0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	57.8	35.6	14.6	61.5	40.3	18.7	62.1	41.1	19.5
k-FWE JPR ($k=1$)	89.8	89.0	88.0	89.3	87.8	84.0	89.8	88.2	85.8
k-FWE JPR ($k=2$)	90.2	89.0	88.9	90.3	88.9	87.2	90.0	89.8	89.4
k-FWE JPR ($k=3$)	89.9	89.5	89.3	90.0	90.2	88.5	90.3	89.3	90.2
Scheffé ($k=1$)	78.2	68.0	54.6	77.8	65.1	47.7	80.1	68.4	51.9
NP Heuristic ($k=1$)	88.2	86.9	84.2	86.3	82.9	75.1	89.1	87.7	84.2
$\boldsymbol{\rho} = -0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	58.6	36.6	15.3	62.2	41.1	19.7	62.6	41.4	19.2
k-FWE JPR ($k=1$)	89.8	89.1	88.6	89.3	87.9	84.0	88.7	87.8	84.9
k-FWE JPR ($k=2$)	89.9	89.4	89.3	90.2	89.4	87.3	89.5	89.3	88.7
k-FWE JPR ($k=3$)	89.7	89.8	89.5	90.7	90.2	88.5	90.4	89.9	89.4
Scheffé ($k=1$)	8.1	0.1	0.0	20.3	5.4	1.0	15.7	2.6	0.1
NP Heuristic ($k=1$)	87.6	85.7	82.3	85.8	82.2	75.0	87.9	85.9	81.1
$\boldsymbol{\rho} = -0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	67.9	55.2	40.7	71.9	60.4	46.4	71.1	58.8	44.1
k-FWE JPR ($k=1$)	90.0	89.8	90.4	89.8	89.3	88.3	89.5	89.4	89.2
k-FWE JPR ($k=2$)	90.0	89.4	89.4	90.2	89.5	87.8	90.0	89.5	89.9
k-FWE JPR ($k=3$)	89.6	89.7	89.3	90.0	89.5	88.0	90.1	89.8	89.8
Scheffé ($k=1$)	0.1	0.0	0.0	0.3	0.1	0.0	0.4	0.1	0.0
NP Heuristic ($k=1$)	87.9	87.6	88.1	87.0	85.9	84.6	87.9	87.3	86.7

Table 3.1: Fixed Lag Length, AR(1), $T = 100$: Empirical $1-k$-FWE Coverage

Nominal Coverage $1 - \alpha_{k\text{-FWE}} = 90\%$

Empirical $1-k$-FWE	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2_3$		
$\boldsymbol{\rho} = 0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	68.5	55.2	38.3	72.2	60.5	44.6	69.8	57.0	40.5
k-FWE JPR ($k=1$)	90.0	90.0	89.9	90.1	90.1	89.8	90.0	90.2	89.8
k-FWE JPR ($k=2$)	89.9	89.9	89.6	90.3	90.2	89.5	90.0	89.9	89.9
k-FWE JPR ($k=3$)	90.0	90.0	89.9	90.1	90.1	89.9	90.1	90.0	90.0
Scheffé ($k=1$)	88.1	88.0	87.9	89.0	88.3	87.5	90.0	89.7	89.4
NP Heuristic ($k=1$)	89.2	88.5	87.9	88.6	87.8	86.5	89.3	88.9	88.5
$\boldsymbol{\rho} = 0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	58.1	34.8	12.8	60.9	38.6	16.0	60.2	37.9	15.3
k-FWE JPR ($k=1$)	89.8	89.8	89.6	90.1	89.7	88.8	89.9	89.7	88.8
k-FWE JPR ($k=2$)	89.9	89.7	89.5	90.5	90.3	89.4	89.8	89.9	89.6
k-FWE JPR ($k=3$)	89.9	89.7	89.9	90.2	90.2	90.0	90.1	90.0	90.1
Scheffé ($k=1$)	80.8	70.5	51.3	80.7	68.3	47.8	81.3	69.3	49.3
NP Heuristic ($k=1$)	89.0	87.7	85.5	88.3	86.4	82.6	89.2	88.1	86.2
$\boldsymbol{\rho} = -0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	57.8	34.8	12.9	61.2	38.8	15.9	61.5	39.6	16.6
k-FWE JPR ($k=1$)	89.9	89.9	88.8	90.0	89.8	89.0	89.8	89.7	88.7
k-FWE JPR ($k=2$)	89.9	89.9	89.8	90.4	89.9	89.6	89.9	89.8	89.4
k-FWE JPR ($k=3$)	90.0	90.1	89.9	90.3	90.1	90.2	90.2	90.1	89.7
Scheffé ($k=1$)	7.1	0.4	0.0	20.0	4.1	0.4	14.0	1.6	0.1
NP Heuristic ($k=1$)	88.5	87.3	84.6	88.3	86.2	82.3	88.7	87.3	85.0
$\boldsymbol{\rho} = -0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	68.4	55.4	38.8	72.6	60.9	45.3	71.4	59.0	42.6
k-FWE JPR ($k=1$)	89.9	90.0	90.4	90.2	90.2	89.8	90.0	90.2	89.7
k-FWE JPR ($k=2$)	89.8	90.1	90.0	90.3	90.3	89.7	89.9	89.9	89.8
k-FWE JPR ($k=3$)	89.9	90.0	89.8	90.0	90.1	90.0	90.0	90.0	89.7
Scheffé ($k=1$)	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0
NP Heuristic ($k=1$)	88.6	87.9	86.8	88.4	87.4	85.8	88.8	87.8	86.8

Table 3.2: Fixed Lag Length, AR(1), $T = 400$: Empirical $1-k$-FWE Coverage

Nominal Coverage $1 - \alpha_{k\text{-FWE}} = 90\%$

Empirical $1-k$-FWE	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2_3$		
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (1.75, -0.85)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	73.6	60.8	45.7	76.9	64.9	50.3	75.7	63.4	49.0
k-FWE JPR ($k=1$)	89.0	88.2	87.2	89.0	87.6	86.6	88.6	88.6	87.2
k-FWE JPR ($k=2$)	88.9	88.3	87.3	89.3	88.0	86.4	88.9	88.7	87.8
k-FWE JPR ($k=3$)	88.8	88.8	88.2	89.5	88.4	86.6	89.3	88.9	88.3
Scheffé ($k=1$)	79.0	73.8	55.8	84.0	79.4	61.2	86.7	82.3	63.8
NP Heuristic ($k=1$)	89.1	90.2	91.4	87.9	87.9	86.8	89.6	90.1	90.4
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (1.25, -0.75)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	63.9	46.9	28.4	69.1	53.9	35.9	68.1	52.3	33.5
k-FWE JPR ($k=1$)	90.0	89.6	89.5	89.4	88.5	86.5	89.3	89.3	88.1
k-FWE JPR ($k=2$)	90.1	89.3	89.3	90.0	88.7	86.7	89.8	89.5	88.7
k-FWE JPR ($k=3$)	89.8	89.6	89.2	90.2	89.6	87.0	90.2	88.8	89.2
Scheffé ($k=1$)	60.7	21.9	5.6	67.9	32.4	13.3	68.9	29.9	10.1
NP Heuristic ($k=1$)	88.6	88.2	87.8	87.1	85.5	81.9	88.7	87.8	86.5
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (-0.65, 0.15)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	63.9	48.2	30.6	67.6	52.9	35.5	67.6	52.1	33.8
k-FWE JPR ($k=1$)	90.1	89.7	89.8	89.5	88.7	86.3	89.1	88.9	87.6
k-FWE JPR ($k=2$)	89.8	89.2	89.2	90.2	89.4	87.5	89.8	89.3	89.2
k-FWE JPR ($k=3$)	89.3	89.4	88.3	90.1	89.3	87.3	89.9	89.6	89.0
Scheffé ($k=1$)	3.6	0.2	0.0	9.7	2.0	0.3	7.2	0.8	0.1
NP Heuristic ($k=1$)	88.1	87.7	87.6	86.4	84.6	81.1	88.2	87.2	85.3
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (-0.7, -0.2)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	60.7	39.9	18.1	64.6	45.2	23.8	64.5	44.6	22.3
k-FWE JPR ($k=1$)	88.8	89.4	88.9	89.2	87.8	84.3	88.8	88.0	85.3
k-FWE JPR ($k=2$)	89.7	89.4	89.4	90.0	88.9	86.0	89.3	88.6	87.4
k-FWE JPR ($k=3$)	89.7	89.7	89.5	90.5	90.0	88.0	90.5	89.9	88.9
Scheffé ($k=1$)	1.8	0.1	0.0	6.9	1.0	0.2	4.1	0.2	0.0
NP Heuristic ($k=1$)	88.2	87.3	85.3	86.2	83.8	78.2	87.9	86.2	82.1

Table 3.3: Fixed Lag Length, AR(2), $T = 100$: Empirical $1-k$-FWE Coverage

Nominal coverage $1 - \alpha_{k\text{-FWE}} = 90\%$

Empirical $1-k$-FWE	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2$		
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (1.75, -0.85)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	75.9	63.2	46.4	78.9	67.5	51.7	76.9	64.5	48.2
k-FWE JPR ($k=1$)	89.8	89.8	89.7	90.2	89.9	89.6	89.9	89.7	89.4
k-FWE JPR ($k=2$)	89.9	89.6	89.5	90.3	90.1	89.3	89.8	89.7	89.4
k-FWE JPR ($k=3$)	90.0	89.7	89.7	90.0	90.0	89.6	90.0	89.9	89.0
Scheffé ($k=1$)	80.2	76.1	56.5	86.2	82.7	65.5	88.4	85.5	89.9
NP Heuristic ($k=1$)	89.2	89.0	88.6	89.1	88.4	87.2	89.3	89.1	88.3
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (1.25, -0.75)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	65.3	47.2	25.8	69.8	53.8	33.5	68.5	51.9	30.9
k-FWE JPR ($k=1$)	89.9	89.9	90.0	90.1	90.1	89.5	89.8	89.9	89.6
k-FWE JPR ($k=2$)	89.9	89.8	89.8	90.3	90.2	89.5	89.8	89.8	89.4
k-FWE JPR ($k=3$)	90.0	89.7	89.7	90.0	90.1	89.6	90.0	90.0	89.8
Scheffé ($k=1$)	62.5	20.0	3.2	72.0	33.2	10.7	72.0	28.4	6.3
NP Heuristic ($k=1$)	88.8	87.9	86.5	88.5	87.4	84.9	89.0	87.9	86.0
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (-0.65, 0.15)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	64.1	47.2	26.6	67.8	52.1	31.8	67.3	51.0	30.6
k-FWE JPR ($k=1$)	89.9	89.7	90.1	90.2	90.1	89.3	89.8	90.0	89.5
k-FWE JPR ($k=2$)	89.9	89.8	89.9	90.4	90.2	89.7	90.0	89.8	89.6
k-FWE JPR ($k=3$)	90.0	89.9	89.7	90.0	90.1	90.0	90.0	89.9	89.8
Scheffé ($k=1$)	3.2	0.1	0.0	9.8	1.3	0.1	6.7	0.5	0.0
NP Heuristic ($k=1$)	88.8	88.0	86.1	88.2	87.0	84.2	88.7	87.7	86.3
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (-0.7, -0.2)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	59.8	37.3	14.8	64.0	42.6	19.2	63.5	42.1	18.9
k-FWE JPR ($k=1$)	89.9	89.8	89.9	89.9	89.9	88.9	89.8	89.8	88.7
k-FWE JPR ($k=2$)	90.1	89.8	89.8	90.1	89.9	89.1	89.9	89.4	89.0
k-FWE JPR ($k=3$)	90.0	90.0	89.8	90.0	90.4	90.1	90.3	90.0	89.6
Scheffé ($k=1$)	1.7	0.1	0.0	6.7	0.5	0.1	3.7	0.1	0.0
NP Heuristic ($k=1$)	88.7	87.8	85.4	88.2	86.6	83.4	88.6	87.3	85.2

Table 3.4: Fixed Lag Length, AR(2), $T = 400$: Empirical $1-k$-FWE Coverage

Nominal coverage $1 - \alpha_{k\text{-FWE}} = 90\%$

Empirical $1-k$-FWE	$T=100$, $\epsilon_t \sim N(0,1)$			$T=400$, $\epsilon_t \sim N(0,1)$		
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (1.75, -0.85)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	72.1	61.8	49.2	76.2	64.5	48.0
k-FWE JPR ($k=1$)	90.4	90.5	89.6	89.8	89.7	87.6
k-FWE JPR ($k=2$)	90.4	89.8	89.7	89.9	89.8	89.7
k-FWE JPR ($k=3$)	90.0	90.3	89.0	90.0	89.7	89.6
Scheffé ($k=1$)	87.9	86.0	64.4	89.2	88.8	66.1
NP Heuristic ($k=1$)	89.2	93.1	95.1	89.8	90.7	90.5
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (1.25, -0.75)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	63.6	46.1	27.0	65.3	47.1	25.5
k-FWE JPR ($k=1$)	90.0	89.4	89.3	89.9	89.8	89.9
k-FWE JPR ($k=2$)	90.2	89.5	89.5	89.9	89.9	89.8
k-FWE JPR ($k=3$)	89.8	89.5	89.3	89.9	89.8	89.7
Scheffé ($k=1$)	63.7	23.2	7.5	66.5	21.6	4.2
NP Heuristic ($k=1$)	87.9	86.7	85.8	88.8	87.8	86.0
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (-0.65, 0.15)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	65.1	48.9	30.4	64.5	47.2	26.2
k-FWE JPR ($k=1$)	90.4	90.1	89.7	90.0	90.0	89.7
k-FWE JPR ($k=2$)	90.5	89.9	89.8	90.1	90.0	90.0
k-FWE JPR ($k=3$)	89.7	89.7	89.6	90.0	89.8	89.8
Scheffé ($k=1$)	2.6	0.2	0.0	2.9	0.1	0.0
NP Heuristic ($k=1$)	88.8	87.9	86.8	89.1	88.0	86.1
$(\boldsymbol{\rho_1}, \boldsymbol{\rho_2}) = (-0.7, -0.2)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals ($k=1$)	59.9	39.5	18.2	59.6	37.3	14.9
k-FWE JPR ($k=1$)	89.4	89.3	88.7	89.9	89.8	89.8
k-FWE JPR ($k=2$)	89.2	89.4	89.8	90.0	90.0	90.0
k-FWE JPR ($k=3$)	89.4	89.7	89.8	90.0	90.1	89.9
Scheffé ($k=1$)	3.0	0.1	0.0	1.9	0.1	0.0
NP Heuristic ($k=1$)	87.8	86.9	85.3	88.7	87.7	85.5

Table 3.5: AR(2), BIC Order Selection: Empirical $1-k$-FWE Coverage

It is worth to pause a moment here and to note that the FWE in the context of multiple testing only depends on the set of true null hypotheses. This is in contrast to definition (3.8) where the FWE depends on all components μ_h. The reasons is that in the context of constructing a joint confidence region for $\boldsymbol{\mu}$, there are no true and false parameters μ_h; they are all 'true' and of interest. Similarly for the definition (3.9) of the FWE in the context of constructing joint prediction regions: all components y_h are 'true' and of interest.

When control of the FWE is deemed too stringent in the context of multiple testing, one can control *generalized error rates* instead. Such generalized error rates are more liberal in terms of rejecting true null hypotheses and, in return, offer a greater ability to reject false null hypotheses.

The most popular generalized error rate, to date, is the *false discovery rate* (FDR). It is the expected value of the *false discovery proportion* (FDP). When applying a multiple testing procedure there will be a (random) total number of R rejections out of the H individual decision problem. Out of these R total rejections, there will be F false rejections (that is, rejections of true null hypotheses). Then one defines

$$\text{FDP} \equiv \frac{F}{R} \quad \text{and} \quad \text{FDR} \equiv \mathbb{E}(\text{FDP}) \,, \tag{3.34}$$

with $\text{FDP} \equiv 0$ in case there are no rejections at all. Control of the FDR amounts to ensuring that $\text{FDR} \leq \gamma$, for some prespecified (small) value $\gamma \in (0, 1)$.

Crucially, the definitions of the FDP and the FDR in the context of multiple testing rely on the notion of a subset of true hypotheses out of the universe of all H hypotheses. But the equivalent of such a subset does not exist in the context of a joint confidence region for $\boldsymbol{\mu}$: all components μ_h are 'true' and of interest. Therefore, controlling the FDR is nonsensical in such a context. In particular, whenever there are any components μ_h at all not contained in the joint confidence region, the FDP is automatically equal to one. And so control of the FDR is actually equivalent to control of the FWE. The reason is that ensuring that $\mathbb{E}(\text{FDP}) \leq \gamma$ is equivalent to ensuring that

$$\mathbb{P}\{\text{At least one } \mu_h \text{ not contained in the JCR}\} \leq \gamma \,. \tag{3.35}$$

For the same reason, control of the FDR is equivalent to control of the FWE also in the context of constructing a joint prediction region for Y.

Remark 3.B.1. As an aside, allow us to point out that the FDR is more popular than it deserves to be. Many applied researchers do not really understand this error rate and the implications when it is applied to a set of data. Since the FDR is the *expected value* of the FDP, little can be said about the realized value of the FDP after applying a multiple testing method which controls the FDR to a set of data. On the other hand, many applied researchers seem to believe that the realized FDP can be at most γ. But such belief is just as valid as believing that the realization of random variable drawn from the standard normal distribution can be at most zero (since the standard normal distribution has expected value zero).

If a statement concerning the realized FDP is the goal, one should control the FDP instead in the sense of ensuring that

$$\mathbb{P}\{\text{FDP} > \gamma\} \leq \alpha \ . \tag{3.36}$$

In this way one can be $1 - \alpha$ confident that the realized FDP is at most γ. The reader is referred to Korn et al. (2004) and Romano et al. (2008) for a more detailed discussion. ■

While control of the FDR is not a meaningful alternative, it is possible to construct joint confidence regions as well as joint prediction regions based on a generalized error rate that is meaningful in these contexts. The solution is to use the *generalized familywise error rate* (k-FWE). Start with the context of multiple testing. For an integer $k \geq 1$, the definition is

$$k\text{-FWE} \ \equiv \mathbb{P}\{\text{Reject at least } k \text{ of the } H_{0,h}\colon\ h \in I(\mathbb{P})\} \ . \tag{3.37}$$

As a special case, the choice $k = 1$ gives back the FWE. On the other hand, any choice $k \geq 2$ results in a less stringent error rate.

Realizing that in the contexts of estimating and forecasting, all components are 'true' and of interest, the definition of the k-FWE can easily be adapted as already described in (3.10)–(3.11). For a joint confidence region (JCR) for $\boldsymbol{\mu}$,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } \mu_h \text{ not contained in the JCR}\} \ ,$$

whereas for a joint prediction region (JPR) for Y,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } y_h \text{ not contained in the JPR}\} \ .$$

Essay 4

Flaws of Scheffe Bands

Why Scheffe Bands are flawed for control of the familiywise error rate (FWE)

Dan Wunderli[1]
Department of Economics, University of Zurich
Wilfriedstrasse 6, CH-8032 Zurich, Switzerland
dan.wunderli@econ.uzh.ch

December 2011

Abstract

This essay explains and proves why Scheffe prediction bands as in Jordà and Marcellino (2010) are flawed from a multiple testing point of view. The critique and proofs are identical for Scheffe confidence bands as in Jordà (2009), only requiring a change in notation. This essay is meant as a potential part or appendix of Essay 3.

[1]Response to Jordà and Marcellino (2010)'s "Path Forecast Evaluation" in Journal of Applied Econometrics and Jordà (2009)'s "Simultaneous Confidence Regions for Impulse Responses" in The Review of Economics and Statistics. This is a working paper version; comments are welcome. Many thanks to Marc Sommer for a useful comment.

4.1 Introduction

The following essay is written for Scheffe prediction bands introduced by Jordà and Marcellino (2010). The critique is similar for Scheffe confidence bands in Jordà (2009); only a change in notation is required.

Let $\widehat{\boldsymbol{Y}}_t(H)$ denote predictions of the random variable $\boldsymbol{y}_t$ over horizons $h = 1, \ldots, H$ as follows

$$\widehat{\boldsymbol{Y}}_t(H) = \begin{pmatrix} \widehat{\boldsymbol{y}}_{t+1} \\ \vdots \\ \widehat{\boldsymbol{y}}_{t+H} \end{pmatrix} \tag{4.1}$$

$\boldsymbol{Y}_t(H) = [\boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_{t+H}]'$ denotes the vector of the H realized future values.

Essay 3 shows how to construct rectangular joint prediction regions (k-FWE Joint Prediction Regions), which are designed to control the probability that k or more realized future values in $\boldsymbol{Y}_t(H)$ lie outside the uniform confidence band at level $\alpha_{k\text{-FWE}}$. With a similar approach, k-FWE joint confidence regions for parameters of the data can be constructed.

This is in contrast to Jordà (2009) and Jordà and Marcellino (2010), who use elliptical joint regions as an approximation to the rectangular joint region. Their method is designed to control what they call a "Wald" error type I; that is, the error type one that all H realized future values in $\boldsymbol{Y}_t(H)$ simultaneously lie outside the uniform prediction band for $\widehat{\boldsymbol{Y}}_t(H)$.

I show why their approach suffers from a severe multiple testing deficiency, leading to degenerate properties with respect to the probability that one or more entries in $\boldsymbol{Y}_t(H)$ lie outside the uniform prediction band, so-called FWE size distortions. Additionally, their theoretically unfounded StepDown modified Scheffe bands are neither joint confidence regions nor joint prediction regions. In this essay, I elaborate on these multiple testing deficiencies and provide closed-forms for the FWE size distortions of these Scheffe bands.

Remark 4.1.1. *The multiple testing deficiency of Scheffe bands holds for joint confidence regions and for joint predictive regions. Thus, I sometimes refer to both sort of regions by the term 'joint regions'. I elaborate on joint prediction regions. For joint confidence regions, the results and proofs are the same, except that null hypotheses exist. Note that* $\mathcal{H}_{0,h} : \boldsymbol{\mu}_h = 0$ *makes sense for joint confidence regions, while '*$\mathcal{H}_{0,h} : \hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h} = 0$*' clearly does not make sense for joint prediction regions. Null hypotheses can only be formulated for parameters such as* $\boldsymbol{\mu}_h$*, not for random variables* $\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}$ *as such.*

4.2 Scheffe bands

Jordà and Marcellino (2010)'s originally proposed Scheffe prediction bands are

$$\widehat{\boldsymbol{Y}}_t(H) \pm P\sqrt{\frac{c_\alpha^2(H)}{H}}\boldsymbol{i}_H \tag{4.2}$$

P is the lower-triangular Cholesky factor of the covariance-matrix Ξ_H/T of $\widehat{\boldsymbol{Y}}_t(H)$, such that $PP' = \Xi_H/T$. $c_\alpha^2(H)$ is the $1-\alpha$ quantile of a χ_H^2 distributed random variable[2], $\boldsymbol{\imath}_H$ is a vector of ones of length H. They assume that $\sqrt{T}(\widehat{\boldsymbol{Y}}_t(H) - \boldsymbol{Y}_t(H) \mid \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots) \xrightarrow{d} N(\boldsymbol{0}, \Xi_H)$ as the sample size $T \to \infty$, which only holds under strong and unrealistic assumptions, as explained in Essay 3. Jordà and Marcellino (2010) propose a so-called StepDown correction to their original Scheffe bands in (4.2) as follows

$$\widehat{\boldsymbol{Y}}_t(H) \pm P \begin{pmatrix} \sqrt{\frac{c_\alpha^2(1)}{1}} \\ \sqrt{\frac{c_\alpha^2(2)}{2}} \\ \vdots \\ \sqrt{\frac{c_\alpha^2(H)}{H}} \end{pmatrix} \tag{4.3}$$

Figure 4.2 and Figure 4.3 show what Scheffe-StepDown prediction bands look like on an AR(1) Monte Carlo data set, as in the simulation setup of Essay 3. The NP heuristic bands, the stringing together of the marginal bootstrap intervals, and the 1-FWE JPR, 2-FWE JPR, and 3-FWE JPR bands are also depicted in these figures. See Essay 3 Section 3.4 for the definitions thereof.

4.3 Multiple Testing Deficiencies of Scheffe bands

Although the original Scheffe prediction bands satisfy the definition of a joint predictive region[3], the StepDown modified Scheffe prediction bands are not joint prediction regions anymore. There is one critical value $\sqrt{c_\alpha^2(H)/H}$ for all H entries in $\widehat{\boldsymbol{Y}}_t(H)$ for the original Scheffe bands. However, in the StepDown modified Scheffe bands, there is a monotonically decreasing sequence of critical values $[\sqrt{c_\alpha^2(h)/h}]_{h=1,\dots,H}$[4].

Furthermore, both the original and the StepDown modified Scheffe bands violate the basic multiple testing insight that $\alpha_{\text{FWE}} \geq \alpha_h$ needs to hold, for the following notation.

$$\text{FWE} = P(\text{Number of points } h \text{ at which } y_{t+h} \text{ lies outside the prediction band} \geq 1) \tag{4.4}$$

where $P(\cdot)$ denotes the probability mechanism. For confidence bands, the FWE is

$$\text{FWE} = P(\text{Number of falsely rejected null hypotheses} \geq 1) \tag{4.5}$$

α_{FWE} is the nominal level of (4.4) or (4.5), α_h denotes the level of each individual prediction or confidence interval. The realized FWE can be much larger than α_{FWE} if one naively uses individual levels $\alpha_h = \alpha_{FWE}$ instead of using proper multiple testing methods. The preceding essays provide ample evidence of this fact.

[2] In principle, it is possible to use bootstrap critical values instead of using the χ^2 distribution, but Jordà and Marcellino (2010) propagate the use of χ^2 critical values

[3] The same holds for the original Scheffe confidence bands and joint confidence regions

[4] As stated in Jordà and Marcellino (2010), page 634

4.3.1 Motivating Examples

Remark 4.3.1. *For the following corollary, the notion of null hypotheses along with the FWE as in* (4.5) *is used for ease of exposition, even though there are no null hypotheses for (Scheffe) prediction bands. Thereafter, the FWE notion* (4.4) *will be used to make it appropriate for Scheffe prediction bands.*

Any sensible multiple testing method ensures that $\alpha_{\text{FWE}} \geq \alpha_h$. This means that a multiple-testing derived critical value c_{α_h} for each individual hypothesis needs to be **larger** in absolute value than the naive individual critical value $c_{\alpha_h=\alpha_{\text{FWE}}}$ not corrected by multiple testing procedures. The following corollary provides an example of this simple fact, which was illustrated for $H = 2$ in Figure 1.1 and Figure 1.2 in the Overview Part I of this Ph.D. thesis.

Corollary 4.3.1. *Consider carrying out $h = 1, \ldots, H$ individual t-tests of $\mathcal{H}_{0,h} : \mu = 0$ at individual level α_h, based on realizations z_h from a standard normally distributed random variable $Z_h \sim N(0,1)$. That is, if the realization z_h is larger than the $1-\alpha$ quantile $c_\alpha^{N(0,1)}$, $\mathcal{H}_{0,h}$ is rejected. The nominal level of the probability of falsely rejecting one or more null hypotheses, denoted as α_{FWE}, is one minus the probability of not rejecting any of the H individual tests*

$$\alpha_{FWE} = 1 - (1 - \alpha_h)^H \tag{4.6}$$

The probability of making one or more false rejections approaches one as the number H of individual t-tests grows to infinity, because α_{FWE} in (4.6) *approaches one as the number of null hypotheses H grows to infinity. In this situation, the multiple testing correction for the individual significance levels α_h, such that the FWE is controlled over the H individual tests, is as follows. By* (4.6) *and $0 < \alpha_h < 1$, we can solve for*

$$\alpha_h = 1 - (1 - \alpha_{FWE})^{1/H} \tag{4.7}$$

This correction (4.7) *of the individual levels is very close to the Bonferroni correction*

$$\alpha_h = \frac{\alpha_{FWE}}{H} \tag{Bonferroni}$$

Note that the multiple-testing corrected individual levels α_h need to be smaller than the desired family-wise level α_{FWE}. As a result, the multiple-testing corrected critical values need to be larger in absolute value than the uncorrected critical value $c_{\alpha_{FWE}}$. In short

$$\alpha_{FWE} \geq \alpha_h \Longrightarrow c_{\alpha_h} \geq c_{\alpha_{FWE}} \tag{4.8}$$

In terms of the familiar standard normal $c_{\alpha/2}^{N(0,1)}$ critical value for a two-sided alternative hypothesis

$$\alpha_1 = 0.1 \geq \alpha_2 = 0.05 \Longrightarrow c_{\alpha_2=0.05/2}^{N(0,1)} = 1.96 \geq c_{\alpha_1=0.1/2}^{N(0,1)} = 1.64 \tag{4.9}$$

In the case of Scheffe bands over $h = 1, \ldots, H$, however, the critical values are reduced, instead of enlarged, in an attempt to "... spread the uncertainty in the orthogonalized path forecast

evenly over all horizons"[5]. Thus, instead of enlarging the region spanned by the marginal intervals, as stipulated by most basic multiple testing insights, Scheffe bands result in a region that is smaller than the rectangular region spanned by the marginal intervals. This leads to large FWE size distortions as is proved in Proposition 4.3.1 and in Proposition 4.3.2. Instead of ensuring that the FWE is smaller or equal than the desired α_{FWE}, Scheffe bands actually increase the FWE to degenerate levels. Before stating and proving these propositions, let us first reproduce and complete Figure 1 in Jordà and Marcellino (2010) as reproduced in Figure 4.1, to see graphically why Scheffe bands are multiple testing deficient. As said in Essay 3, using asymptotic normality for predictions only makes sense under strong, unrealistic assumptions; asymptotic normality for confidence intervals or regions holds under milder conditions.

Nonetheless, by using normality, in the top panel of Figure 4.1[6], the naive joint prediction region spanned by the marginal two-sided prediction intervals at level $\alpha = 0.05$ is $[-1.96, 1.96] \times [-1.96, 1.96]$, depicted as a solid red rectangle. Under normality, any sensible multiple-testing correction to the individual critical values 1.96 leads to an enlargement of 1.96, thus ensuring that the realized FWE as in (4.4) is smaller or equal than the desired α_{FWE}.

Any sensible multiple-testing corrected rectangular region needs to be **larger** than $[-1.96, 1.96] \times [-1.96, 1.96]$. Nonetheless, Scheffe bands result in a joint rectangular region $[-1.73, 1.73] \times [-1.73, 1.73]$, depicted as a dashed rectangle, that is **smaller** than the marginal $[-1.96, 1.96] \times [-1.96, 1.96]$. Instead of lowering the FWE committed by the marginal intervals, Scheffe bands increase this error type (FWE) vis-à-vis the marginal intervals. The elliptical joint region of the orthogonalized predictions are depicted as a dashed circle, which Scheffe bands use to approximate the rectangular joint region.

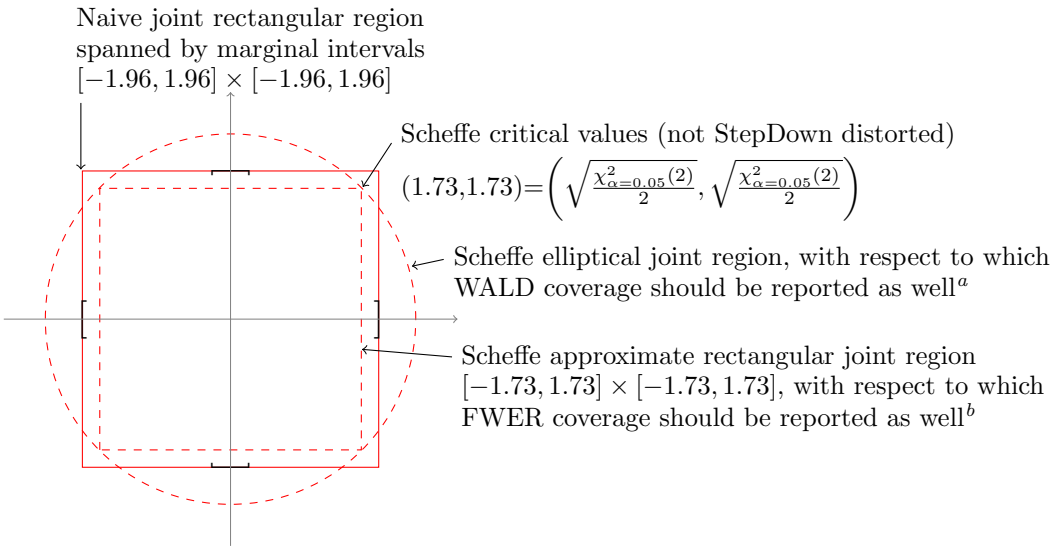
Remark 4.3.2. *Note that* $\sqrt{\frac{c_\alpha^2(1)}{1}} = c_{\alpha/2}^{N(0,1)}$ *holds. For* $H \geq 2$, $\sqrt{\frac{c_\alpha^2(H)}{H}} < c_{\alpha/2}^{N(0,1)}$ *holds for* $\alpha \leq \tilde{\alpha}(H)$ *because for* $X_H \sim \chi_H^2$, $X_H \sim N(H, 2H)$ *holds for large* H [7]. *Thus, the example above that the Scheffe rectangular region* $[-1.73, 1.73] \times [-1.73, 1.73]$ *is smaller than the* $N(0,1)$ *derived one* $[-1.96, 1.96] \times [-1.96, 1.96]$ *does not generalize to all* $0 < \alpha < 1$ *for all* H. *Numerical evidence shows that* $\sqrt{\frac{c_\alpha^2(H)}{H}} < c_{\alpha/2}^{N(0,1)}$ *holds for* $\alpha \leq \tilde{\alpha}(H) = 0.2$ *for any* $H \geq 2$. *For the propositions proved below, however, how* $\sqrt{\frac{c_\alpha^2(H)}{H}}$ *compares to* $c_{\alpha/2}^{N(0,1)}$ *is not an issue.*

The bottom panel of Figure 4.1 depicts the fact that any multiple-testing corrected rectangular joint region needs to be larger than the region spanned by the marginal intervals, which is $[-1.96, 1.96] \times [-1.96, 1.96]$ under asymptotic normality. Additionally, the bottom panel of Figure 4.1 depicts that the Scheffe rectangular joint region degenerates to a hyper-cube $[-1, 1] \times \ldots \times [-1, 1]$ as $H \to \infty$, as shown in the proofs below. It is clear that this is a degenerate asymptotic property; conceptually, a rectangular joint region at level $1 - \alpha_{\text{FWE}}$ $(0 < \alpha_{\text{FWE}} < 1)$ needs to converge to the open hyper-cube $(-\infty, \infty) \times \ldots \times (-\infty, \infty)$ as $H \to \infty$ if one wanted to control the FWE at level α_{FWE} in this conceptual situation.

[5] Jordà and Marcellino (2010), p. 641, second paragraph

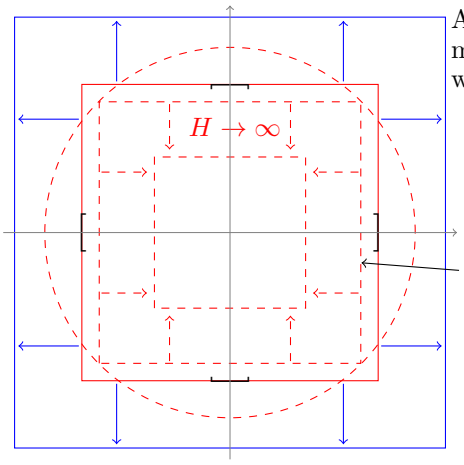
[6] Corresponding to Figure 1 in Jordà and Marcellino (2010)

[7] Proof thereof in proof to Proposition 4.3.1(b) below



[a]Jordà and Marcellino (2010) report WALD coverage with respect to the StepDown modified Scheffe elliptical joint region, although Jordà (2009) reports with respect to the original Scheffe bands

[b]Jordà and Marcellino (2010) report FWER coverage with respect to the StepDown modified Scheffe approximate rectangular joint region, although Jordà (2009) reports with respect to the original Scheffe bands



Any sensible multiple testing corrected joint region
must be LARGER or equal to product of marginal intervals,
which is implied by the general $\alpha_h \leq \alpha_{\text{FWE}} \Rightarrow c_{\alpha_h} \geq c_{\alpha_{\text{FWE}}}$[a].

Scheffe approximate joint rectangular region is SMALLER
than "1 S.E. rectangle" for $\alpha < \tilde{\alpha}(H)$, see Remark 4.3.2
As $H \to \infty$, the Scheffe approximate rectangular joint region
degenerates to the hyper-cube $[-1, 1] \times \cdots \times [-1, 1]$
with realized FWE = 1 instead of realized FWE $\simeq \alpha_{\text{FWE}}$,
see Proposition 4.3.1 and 4.3.2

[a]As in equation (4.8)

Figure 4.1: Why Scheffe bands are multiple testing deficient

4.3.2 Starting point independent entries

With these graphs as a helping guide in mind, I start off with deriving the FWE size distortion as in (4.4) of Scheffe prediction bands in the case of i.i.d. entries of $\widehat{\boldsymbol{Y}}_t(H)$. From these results, the results for the more general case follow.

An FWE size distortion is the difference between the nominal significance level α_{FWE} and the realized FWE. If not available in closed-form, the realized FWE can be simulated as in Essay 3. In using the normal distribution as for Scheffe bands, however, one knows the probability of committing a family-wise error in closed-form, at least asymptotically as the size of the data set goes to infinity.

Proposition 4.3.1 (FWE Size Distortion of Scheffe bands for i.i.d. entries of $\widehat{\boldsymbol{Y}}_t(H)$)**.** *Consider H independent draws from the standard normal distribution, the realizations of which are stacked in the H dimensional vector $\widehat{\boldsymbol{Y}}_t(H)$. Then the following holds for a desired nominal level $0 < \alpha_{FWE} < 1$*

(a) the FWE size distortion of the original Scheffe bands for $\widehat{\boldsymbol{Y}}_t(H)$ is

$$\left(1 - \prod_{h=1}^{H} \Phi\left(\sqrt{\frac{c_\alpha^2(H)}{H}}\right)\right) - \alpha_{FWE}.$$

$\Phi : \mathbb{R} \to [0,1]$ denotes the cumulative distribution function of the standard normal distribution, $\alpha = \alpha_{FWE}$ is chosen in $c_\alpha^2(H)$.

(b) the FWE size distortion approaches $1 - \alpha_{FWE}$ as $H \to \infty$.

(c) the StepDown modification of the Scheffe bands eases the finite-sample FWE size distortion, but these Scheffe bands are not joint regions anymore and are larger than without the StepDown modification.

(d) the FWE size distortion of the StepDown modified Scheffe bands grows to $1 - \alpha_{FWE}$ as $H \to \infty$.

Proof. The Scheffe bands for $\widehat{\boldsymbol{Y}}_t(H)$ are

$$\widehat{\boldsymbol{Y}}_t(H) \pm P\sqrt{\frac{c_\alpha^2(H)}{H}}\boldsymbol{i}_H \tag{4.10}$$

Since the independent draws are from the standard normal distribution, the covariance matrix of $\widehat{\boldsymbol{Y}}_t(H)$ is $\Xi_H/T = \mathbf{I}_H$ and its Cholesky factor is $P = \mathbf{I}_H$. Hence, the Scheffe band can be rewritten as

$$\widehat{\boldsymbol{Y}}_t(H) \pm \sqrt{\frac{c_\alpha^2(H)}{H}}\boldsymbol{i}_H \tag{4.11}$$

(a) Committing a family-wise error type I (FWE) here means committing an error type one at $h = 1$ or at $h = 2$ or ... or at $h = H$. The probability of this event is one minus

the probability of not committing an error type one at any $h = 1, \ldots, H$. As $c_\alpha^2(H)$ is a population quantile, not a random variable, the FWE is one minus the product over $h = 1, \ldots, H$ of the probability of not committing an error type one at h. The probability of not committing an error type one at h is $\Phi\left(\sqrt{c_\alpha^2(H)/H}\right)$. The FWE size distortion of the Scheffe bands (without StepDown modification) is the difference between the realized FWE and the nominal α_{FWE}

$$\Delta\text{SizeScheffe}(H) = \left(1 - \prod_{h=1}^{H} \Phi\left(\sqrt{\frac{c_\alpha^2(H)}{H}}\right)\right) - \alpha_{\text{FWE}} \tag{4.12}$$

where $\alpha = \alpha_{\text{FWE}}$ is chosen in the Scheffe critical value $c_\alpha^2(H)$.

(b) It can be proved that the Scheffe critical value

$$\sqrt{\frac{c_\alpha^2(H)}{H}} \to 1 \quad \text{as } H \to \infty \tag{4.13}$$

as follows. A χ^2 distributed random variable $X_k = \sum_{i=1}^{k} Z_i \sim \chi_k^2$ is a sum of k independent random variables $Z_i \sim N(0,1)$ with finite mean and variance. The mean of X_k is k, its variance it $2k$. By $\frac{X_k - k}{\sqrt{2k}} \xrightarrow{d} N(0,1)$, it holds that $X_k \sim N(k, 2k)$ for large k. It follows that $X_k/k \sim N(k/k, 2k/k^2) = N(1, 2/k)$ for large k. This means that $X_k/k \xrightarrow{p} 1$, because the variance $2/k$ of X_k/k vanishes as the degrees of freedom $k \to \infty$. Formally, by the Chebyshev inequality, it holds or arbitrary $\epsilon > 0$ that

$$0 \leq P\left(\left\{\left|\frac{X_k}{k} - 1\right| > \epsilon\right\}\right) \leq \frac{\text{Var}(X_k/k)}{\epsilon} = \frac{2}{\epsilon k} \tag{4.14}$$

from which

$$\lim_{k\to\infty} P\left(\left\{\left|\frac{X_k}{k} - 1\right| > \epsilon\right\}\right) = 0 \quad \Longleftrightarrow \frac{X_k}{k} \xrightarrow{p} 1 \quad \text{as } k \to \infty \tag{4.15}$$

follows. Now,

$$\begin{aligned}\frac{c_\alpha^2(H)}{H} = \frac{F_H^{-1}(1-\alpha)}{H} = \frac{\inf\{X_H : F_H(X_H) \leq 1-\alpha\}}{H} \\ = \inf\left\{\frac{X_H}{H} : F_H\left(\frac{X_H}{H}\right) \leq 1-\alpha\right\}\end{aligned} \tag{4.16}$$

By (4.15), the asymptotic cumulative distribution function $F(x)$ of X_H/H is

$$F(x) = \begin{cases} 1 & x = 1 \\ 0 & \text{else} \end{cases}. \tag{4.17}$$

Hence, the set within the $\inf\{.\}$ from the most right-hand term in (4.16) converges as follows

$$\left\{\frac{X_H}{H} : F_H\left(\frac{X_H}{H}\right) \leq 1-\alpha\right\} \xrightarrow{H\to\infty} \{x : F(x) \leq 1-\alpha\} = \{x : x \neq 1\} = \{x : x > 1\}, \tag{4.18}$$

for $0 < \alpha < 1$ by $X_H/H \xrightarrow{p} 1$ and X_H/H being monotonically decreasing in H [8], from which

$$\inf\left\{\frac{X_H}{H} : F_H\left(\frac{X_H}{H}\right) \leq 1-\alpha\right\} \overset{H\to\infty}{\longrightarrow} \inf\{x : x > 1\} = 1 \tag{4.19}$$

follows. Thus, any α-quantile of $X_H/H \xrightarrow{p} 1$ $(0 < \alpha < 1)$ as in (4.16) attains the value one as $k \to \infty$. We have thus proved that for $0 < \alpha < 1$

$$\frac{c_\alpha^2(H)}{H} \to 1, \tag{4.20}$$

from which $\sqrt{\frac{c_\alpha^2(H)}{H}} \to 1$ follows. Note that this is a deterministic convergence, since $c_\alpha^2(H)$ is a population quantile, not a random variable.

Hence, the FWE size distortion of the original Scheffe bands converges as follows

$$\Delta\text{SizeScheffe}(H) \to \left(1 - \prod_{h=1}^{\infty} \Phi(1)\right) - \alpha_{\text{FWE}} = 1 - \alpha_{\text{FWE}} \quad \text{as } H \to \infty \tag{4.21}$$

since $0 < \Phi(1) < 1$.

(c) The StepDown modification of the Scheffe bands consists in using the monotonically decreasing sequence of critical values

$$\left(\sqrt{\tfrac{c_\alpha^2(1)}{1}}, \quad \ldots, \quad \sqrt{\tfrac{c_\alpha^2(H)}{H}}\right) \tag{4.22}$$

instead of using one and the same critical value $\sqrt{c_\alpha^2(H)/H}$ for all H entries in $\widehat{\boldsymbol{Y}}_t(H)$. Hence,

$$\Delta\text{SizeScheffeStepDown}(H) = \left(1 - \prod_{h=1}^{H} \Phi\left(\sqrt{\frac{c_\alpha^2(h)}{h}}\right)\right) - \alpha_{\text{FWE}} \tag{4.23}$$

As $\sqrt{c_\alpha^2(h)/h} > \sqrt{c_\alpha^2(H)/H}$ for $h < H$, the FWE size distortion of the StepDown modified Scheffe bands is mitigated as follows. If an error type one occurs at $h = \tilde{h} < H$ with respect to the original critical value $\sqrt{c_\alpha^2(H)/H}$, the larger StepDown modified critical value $\sqrt{c_\alpha^2(\tilde{h})/\tilde{h}}$ may not commit this error type one at $h = \tilde{h}$. Thus, at each h, the probability of not committing an error type one is larger than before the StepDown modification: $\Phi\left(\sqrt{c_\alpha^2(h)/h}\right) > \Phi\left(\sqrt{c_\alpha^2(H)/H}\right)$. By aforementioned monotonicity properties,

$$\Delta\text{SizeScheffeStepDown}(H) < \Delta\text{SizeScheffe}(H). \tag{4.24}$$

(d) Since each factor $\Phi\left(\sqrt{\frac{c_\alpha^2(h)}{h}}\right) < 1$,

$$\Delta\text{SizeScheffeStepDown}(H) \to 1 - \alpha_{\text{FWE}} \quad \text{as } H \to \infty \tag{4.25}$$

follows.

■

[8] As stated in Jordà and Marcellino (2010), page 634

4.3.3 Non-negatively correlated entries

It seems to be clear that these FWE size distortions do not magically vanish in the general case of Scheffe bands for $\widehat{\boldsymbol{Y}}_t(H)$, where the H entries are not identically distributed, and correlated. Informally, one may think that the Cholesky factor P in the Scheffe band projects the critical values $\sqrt{c_\alpha^2(H)/H}$ from the uncorrelated χ^2_α onto the correlated $\widehat{\boldsymbol{Y}}_t(H)$. This seems to be parallel to generating a random vector $\widehat{\boldsymbol{Y}}_t(H)$ with covariance matrix Ξ_H by projecting a generated i.i.d. vector $\boldsymbol{X}_H$ onto the Ξ_H-correlated world by using the Cholesky factor P of Ξ_H as $\widehat{\boldsymbol{Y}}_t(H) = P\boldsymbol{X}_H$. However, the critical values are not correlated; population quantiles of the χ^2 distribution are fixed numbers, thus cannot be correlated. That being said, let us alter Proposition 4.3.1 above for the case of non-negatively correlated entries in $\widehat{\boldsymbol{Y}}_t(H)$.

Proposition 4.3.2 (FWE size distortion of Scheffe bands for non-negatively correlated entries of $\widehat{\boldsymbol{Y}}_t(H)$)**.** *Let $\widehat{\boldsymbol{Y}}_t(H)$ be a vector of length H with non-negatively correlated entries, generated from a stationary random process. Then, the following holds for a nominal level α_{FWE}*

(a) The FWE size distortion of Scheffe and Scheffe-StepDown bands is strictly monotonically increasing in H.

(b) The FWE size distortion of Scheffe and Scheffe-StepDown bands approaches $1-\alpha_{FWE}$ as $H \to \infty$.

(c) The FWE size distortion of Scheffe and Scheffe-StepDown bands is strictly monotonically decreasing in the strength of autocorrelation in $\widehat{\boldsymbol{Y}}_t(H)$ if $\boldsymbol{Y}_t(H)$ follows a stationary AR(1) process.

Proof. One can write the lower-triangular Cholesky factor P of Ξ_H/T as follows

$$P = \begin{pmatrix} p_{11} & 0 & \dots & 0 \\ p_{21} & p_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p_{H1} & p_{H2} & \dots & p_{HH} \end{pmatrix} = \begin{pmatrix} p_{1.} \\ p_{2.} \\ \vdots \\ p_{H.} \end{pmatrix}. \tag{4.26}$$

The Scheffe and the Scheffe-StepDown prediction bands can be written as follows, respectively,

$$\text{Scheffe: } \widehat{\boldsymbol{Y}}_t(H) \pm \begin{pmatrix} p_{1.}\mathbf{1}_H\sqrt{\frac{c_\alpha^2(H)}{H}} \\ p_{2.}\mathbf{1}_H\sqrt{\frac{c_\alpha^2(H)}{H}} \\ \vdots \\ p_{H.}\mathbf{1}_H\sqrt{\frac{c_\alpha^2(H)}{H}} \end{pmatrix}, \tag{4.27}$$

$$\text{SheffeStepDown: } \widehat{\boldsymbol{Y}}_t(H) \pm \begin{pmatrix} p_{1.}\left(\sqrt{\frac{c_\alpha^2(1)}{1}}, \dots, \sqrt{\frac{c_\alpha^2(H)}{H}}\right)' \\ p_{2.}\left(\sqrt{\frac{c_\alpha^2(1)}{1}}, \dots, \sqrt{\frac{c_\alpha^2(H)}{H}}\right)' \\ \vdots \\ p_{H.}\left(\sqrt{\frac{c_\alpha^2(1)}{1}}, \dots, \sqrt{\frac{c_\alpha^2(H)}{H}}\right)' \end{pmatrix}. \tag{4.28}$$

Since $p_{hj} = 0$ for $j > h$ by definition of a lower-triangular matrix,

$$\text{ScheffeStepDown: } \widehat{\boldsymbol{Y}}_t(H) \pm \begin{pmatrix} p_{11}\sqrt{\frac{c_\alpha^2(1)}{1}} \\ p_{21}\sqrt{\frac{c_\alpha^2(1)}{1}} + p_{22}\sqrt{\frac{c_\alpha^2(2)}{2}} \\ \vdots \\ p_{H1}\sqrt{\frac{c_\alpha^2(1)}{1}} + \cdots + p_{HH}\sqrt{\frac{c_\alpha^2(H)}{H}} \end{pmatrix}. \tag{4.29}$$

The FWE is still one minus the product of not committing an error type one at each $h = 1, \ldots, H$; the population χ^2 critical values are not random, thus cannot be correlated. As a result,

$$\Delta\text{SizeScheffe}(H) = \left(1 - \prod_{h=1}^{H} \Phi\left(p_{h\cdot}\mathbf{1}_H\sqrt{\frac{c_\alpha^2(H)}{H}}\right)\right) - \alpha_{\text{FWE}} \tag{4.30}$$

$$\Delta\text{SizeScheffeStepDown}(H) = \left(1 - \prod_{h=1}^{H} \Phi\left(p_{h\cdot}\left(\sqrt{\frac{c_\alpha^2(1)}{1}}, \ldots, \sqrt{\frac{c_\alpha^2(H)}{H}}\right)'\right)\right) - \alpha_{\text{FWE}} \tag{4.31}$$

(a) For positively correlated entries in $\widehat{\boldsymbol{Y}}_t(H)$, $p_{h\cdot} \geq 0$ holds by Lemma 4.A.1. Thus, for the original Scheffe bands

$$\prod_{h=1}^{H} \Phi\left(p_{h\cdot}\mathbf{1}_H\sqrt{\frac{c_\alpha^2(H)}{H}}\right) > \prod_{h=1}^{H+1} \Phi\left(p_{h\cdot}\mathbf{1}_{H+1}\sqrt{\frac{c_\alpha^2(H+1)}{H+1}}\right) \tag{4.32}$$

because each factor is smaller than one. As $\sqrt{c_\alpha^2(h)/h} > \sqrt{c_\alpha^2(H)/H}$ for $h < H$, the same holds for the Scheffe-StepDown bands

$$\prod_{h=1}^{H} \Phi\left(p_{h\cdot}\left(\sqrt{\frac{c_\alpha^2(1)}{1}}, \ldots, \sqrt{\frac{c_\alpha^2(H)}{H}}\right)'\right) > \prod_{h=1}^{H+1} \Phi\left(p_{h\cdot}\left(\sqrt{\frac{c_\alpha^2(1)}{1}}, \ldots, \sqrt{\frac{c_\alpha^2(H+1)}{H+1}}\right)'\right) \tag{4.33}$$

For uncorrelated entries, the proof that the strict inequalities above hold is as in the proof of Proposition 4.3.1(a), except that the Cholesky factor P is not a diagonal matrix with ones on the diagonal anymore. Hence,

$$\Delta\text{SizeScheffe}(H) < \Delta\text{SizeScheffe}(H+1) \tag{4.34}$$

$$\Delta\text{SizeScheffeStepDown}(H) < \Delta\text{SizeScheffeStepDown}(H+1) \tag{4.35}$$

for non-negatively correlated entries in $\widehat{\boldsymbol{Y}}_t(H)$.

(b) Follows from (4.30), (4.31), Proposition 4.3.2(a), and Lemma 4.A.1.

(c) Since the process is non-negatively correlated, all entries of the covariance matrix Ξ_H/T are non-negative. Thus, all entries of its Cholesky factor P are non-negative by Lemma 4.A.1. From (a), the probability of not committing an error type one at any $h = 1, \ldots, H$

$$\prod_{h=1}^{H} \Phi\left(p_{h\cdot}\mathbf{1}_H\sqrt{\frac{c_\alpha^2(H)}{H}}\right) \to 0 \quad \text{as } H \to \infty \tag{4.36}$$

The speed of convergence is faster the lower the non-negative correlation of the process $\widehat{\boldsymbol{Y}}_t(H)$ is. Think of two AR(1) processes with $\rho_1 = 0.9$ versus one with $\rho_2 = 0.5$. It can be shown that for any $h \leq H$, the following holds $p_{h.}\mathbf{1}_H \mid \rho_1 > p_{h.}\mathbf{1}_H \mid \rho_2$ for $\rho_1 > \rho_2$. Thus, for a given H, in the case of two AR(1) processes with $\rho_1 > \rho_2$

$$\prod_{h=1}^{H} \Phi\left(p_{h.}\mathbf{1}_H \sqrt{\frac{c_\alpha^2(H)}{H}}\right) \mid \rho_1 > \prod_{h=1}^{H} \Phi\left(p_{h.}\mathbf{1}_H \sqrt{\frac{c_\alpha^2(H)}{H}}\right) \mid \rho_2 \tag{4.37}$$

Hence,

$$\Delta\text{SizeScheffe}(H) \mid \rho_1 < \Delta\text{SizeScheffe}(H) \mid \rho_2 \tag{4.38}$$

and

$$\Delta\text{SizeScheffeStepDown}(H) \mid \rho_1 < \Delta\text{SizeScheffeStepDown}(H) \mid \rho_2 \tag{4.39}$$

for two AR(1) processes with $\rho_1 > \rho_2$.

■

Remark 4.3.3. *I do not know yet whether I can extend the proofs to negatively correlated entries in $\widehat{\boldsymbol{Y}}_t(H)$ as well. For an AR(1) process, it can be shown that $|p_{h.}\mathbf{1}_H \mid \rho_1| > |p_{h.}\mathbf{1}_H \mid \rho_2|$ for $|\rho_1| > |\rho_2|$. But the probability of not committing an error type one in the proofs above depends on $p_{h.}$, not on the absolute value thereof. Thus, it is not obvious if inequalities such as* (4.37) *hold for negatively correlated entries in $\widehat{\boldsymbol{Y}}_t(H)$ as well. Additionally, maybe I should provide proofs for AR(p) or VAR(p) processes as well here, by suitably redefining 'strength of autocorrelation' and ensuring that the monotonicity property in Lemma 4.A.1 holds.*

4.4 Discussion of Simulation Results

Both deficient properties in Proposition 4.3.2(a) and 4.3.2(c) are reported by means of an AR(1) simulation in Jordà and Marcellino (2010) as well as in the numerical study of Essay 3. This latter numerical evidence shows that property 4.3.2(c) extends to negatively correlated AR(1) processes: The FWE size distortion for $\boldsymbol{\rho} = -0.5$ is smaller than the one for $\boldsymbol{\rho} = -0.9$, as can be seen in Table 3.1 and in Table 3.2. Note that these Tables in Essay 3 report the $1 -$ FWE coverage of Scheffe bands.

The same property 4.3.2(c) holds for AR(2) processes with coefficients (ρ_1, ρ_2), where strength of autocorrelation is defined as $\rho_1 + \rho_2$, as shown in Table 3.3 and in Table 3.4.

The monotonically increasing FWE size distortion of Scheffe bands for increasing H, property 4.3.2(a), is evident in the numerical results of Essay 3 by comparing the FWE coverage for increasing $H \in \{6, 12, 24\}$.

Additionally, the FWE in the AR(1) and AR(2) cases of negative autocorrelation is close to or equal to one for $H = 12$ and $H = 24$, which may be a remarkable indication for the $H \to \infty$ asymptotic property 4.3.2(b) holding for $H = 24$ already in some cases.

Furthermore, in light of the results proved here, the FWE size distortions reported in Jordà (2009) and Jordà and Marcellino (2010) seem unduly small in their reported Stock and Watson (2001) vector autoregression (VAR) Monte Carlo study. This may be due to strong autocorrelation in the three variables of the Stock and Watson (2001) VAR; in the AR(1) case, strong autocorrelation lowers the FWE size distortion by 4.3.2(c). The propositions for VAR(p) processes, however, are not part of this essay.

4.5 Conclusions

Scheffe and Scheffe-StepDown bands violate basic multiple testing insights, thus they shall not be used if one cares about any sort of multiple error type one. The StepDown modified Scheffe bands are neither joint confidence regions nor joint predictive regions anymore, hence they additionally violate another basic statistical concept. Furthermore, this StepDown modification partly covers up the serious FWE size deficiencies that the Scheffe bands exhibit.

In the AR(1) simulations of Jordà and Marcellino (2010) and in the AR(1) and AR(2) simulations of Essay 3, the FWE deficient properties of Scheffe bands that were proved in this essay are evident.

That the FWE size distortions do not show up clearly in the empirical application of Jordà (2009) and Jordà and Marcellino (2010) may be due to strong autocorrelation of the variables within the Stock and Watson (2001) vector autoregression. I proved in this essay that for an AR(1) process, the stronger the positive autocorrelation is, the lower is the FWE size distortion of Scheffe and Scheffe-StepDown bands.

If one cares about the error type one with respect to an elliptic joint region, the Scheffe bands without the StepDown modification may provide reasonable control of the associated error type one, called Wald error type one in Jordà (2009). Whether control of this error type one makes more sense than FWE control must be decided by the careful practitioner. If the concern is that the FWE is overly strict, control of the k-FWE as in Essay 3 offers a viable solution.

References

Please find the references at the end of this Ph.D. thesis.

4.A Entrywise Monotonicity Property of Cholesky Factor

The aim of this appendix is to prove that the lower-triangular Cholesky factor L of a matrix A with positive entries has positive lower-triangular entries $L_{ij} > 0$ $(i > j)$.

Lemma 4.A.1. *Let A denote a symmetric, positive definite matrix with positive entries $A_{ij} > 0$, and let L denote its lower-triangular Cholesky factor. The entries A_{ij} decrease monotonically as one moves further away from the diagonal in the lower-triangular part of A: $A_{ij} > A_{i(j+k)}$ and $A_{ij} > A_{(i+k)j}$ for $i > j$, $k > 0$. Then, all lower-triangular entries $\{L_{ij}, i > j\}$ are also strictly positive: $L_{ij} > 0$ for $i > j$.*

The covariance matrix A of a positively autocorrelated AR(1) process satisfies the monotonicity requirements of Lemma 4.A.1. The main line of the following proof is that for $L_{ij} < 0$ $(i > j)$ to hold, two observations x_t, x_{t+k} $(k > 0)$ need to be more strongly autocorrelated as k increases, which is a contradiction for a stationary, positively autocorrelated AR(1) process. This contradiction shows up in the proof below by violation of the monotonicity $A_{ij} > A_{i(j+k)}$ and $A_{ij} > A_{(i+k)j}$ for $i > j$, $k > 0$.

Proof. In the Cholesky-Banachiewicz algorithm, the diagonal entries of the Cholesky factor L are given by

$$L_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2} \quad > 0, \tag{4.40}$$

while the off-diagonal entries of L are given by

$$L_{ij} = \frac{1}{L_{jj}} \left(A_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} \right) \text{ for } i > j, \tag{4.41}$$

$$L_{ij} = 0 \text{ for } i < j. \tag{4.42}$$

That $L_{jj} > 0$ holds is a well-known property, it can be found in Bhatia (2007) 1.1.(iv)[9]. What needs to be proved here is that $L_{ij} > 0$ for $i > j$. I do so with a proof by contradiction. I will suppose that $L_{ij} < 0$ holds for some i, j, which will violate the monotonicity $A_{ij} > A_{i(j+k)}$ and $A_{ij} > A_{(i+k)j}$ for $i > j$, $k > 0$.

The Cholesky-Banachiewicz algorithm starts off in the upper left corner of L and progresses row by row. Bear in mind that all entries of A are positive. We know by (4.40) that $L_{11} = \sqrt{A_{11}} > 0$; $L_{1j} = 0$ for $j > 1$ holds by L being lower-triangular. Furthermore,

$$L_{21} = \frac{1}{L_{11}} A_{21} > 0 \tag{4.43}$$

[9] Bhatia (2007) refers to positive semidefinite matrices by 'positive matrices'. Thus, the statement that the diagonal of a Cholesky factor of a positive matrix can be chosen as nonnegative applies to the diagonal of the Cholesky factor L of a positive semidefinite matrix A.

holds by $L_{11} > 0$ and $A_{ij} > 0$ for all i, j. $L_{22} = \sqrt{A_{22} - L_{21}^2} > 0$ by (4.40). $L_{2j} = 0$ for $j > 2$ holds by L being lower-triangular. Then,

$$L_{31} = \frac{1}{L_{11}} A_{31} > 0 \tag{4.44}$$

follows trivially.

Assume for the sake of argument that $L_{32} < 0$ holds, where

$$L_{32} = \frac{1}{L_{22}}(A_{32} - L_{31}L_{21}) = \frac{A_{32} - \frac{A_{31}A_{21}}{L_{11}^2}}{L_{22}} = \frac{L_{11}^2 A_{32} - A_{31}A_{21}}{L_{11}^2 L_{22}} = \frac{A_{11}A_{32} - A_{31}A_{21}}{A_{11}L_{22}} \tag{4.45}$$

The denominator of (4.45) is positive by (4.40) and $A_{ij} > 0 \; \forall i, j$. Thus, the nominator of (4.45) needs to be negative to satisfy the assumption $L_{32} < 0$. This is in contradiction to the monotonicity $A_{ij} > A_{i(j+k)}$ and $A_{ij} > A_{(i+k)j}$ for $i > j$, $k > 0$, which is most evident for A being the covariance matrix of an AR(1) process. Since $A_{32} = A_{21}$, $A_{11}A_{32} < A_{31}A_{21} \Leftrightarrow A_{11} < A_{31}$ needs to hold, which violates the monotonicity $A_{ij} > A_{i(j+k)}$ and $A_{ij} > A_{(i+k)j}$ for $i > j$, $k > 0$.

The proof by contradiction for any L_{ij} $(i > j)$ is likewise: Assume that $L_{ij} < 0$, use that all entries of L that have already been calculated by the Cholesky-Banachiewicz algorithm are larger than zero, show that $L_{ij} < 0$ can only hold if the monotonicity in A as described in Lemma 4.A.1 is violated. The details are left to the reader. ■

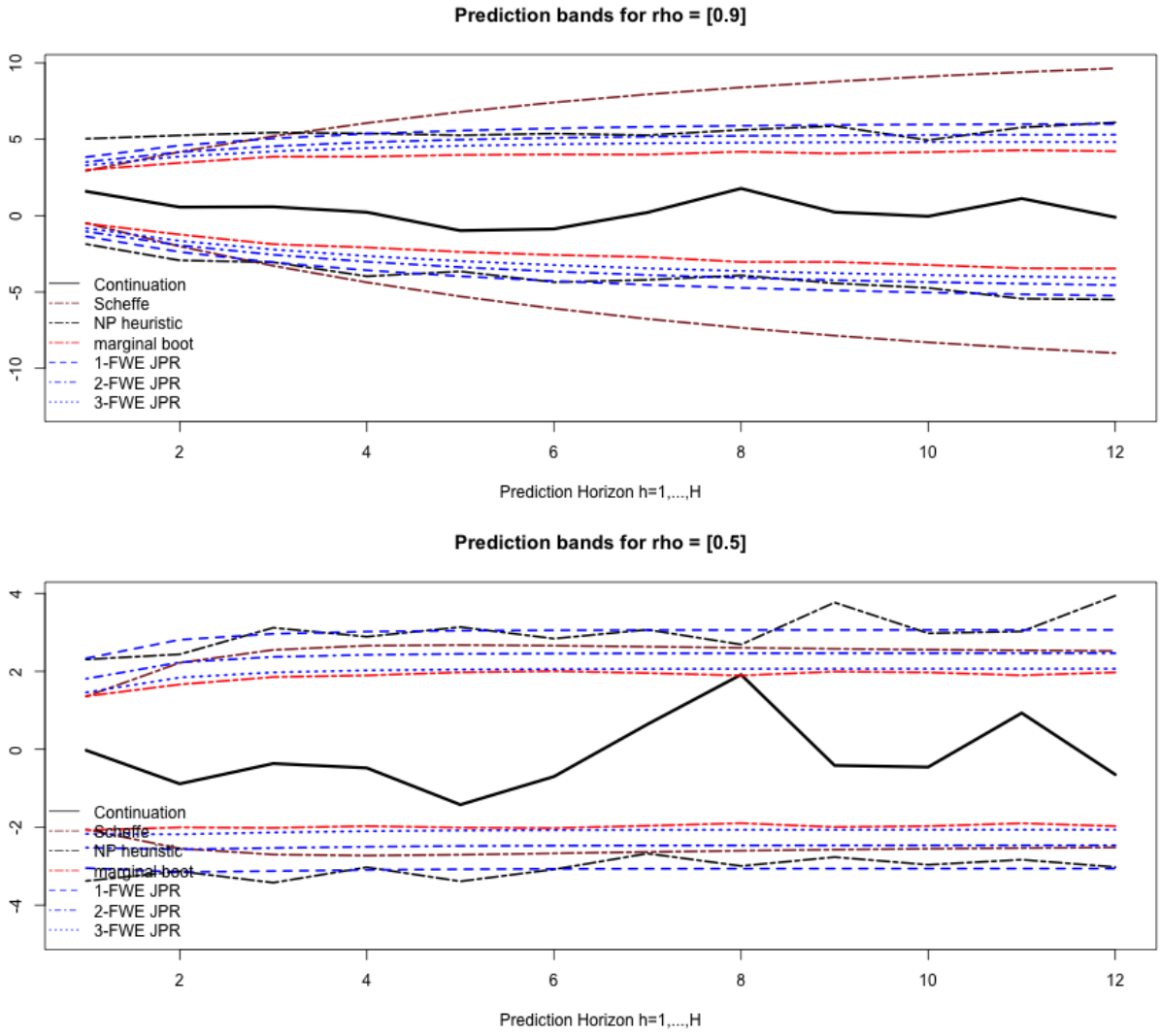


Figure 4.2: AR(1) forecast bands $FCB(\widehat{y}^{(m)}_{H=12})$ and first continuation $y^{(m),(c=1)}_{(T+1):(T+12)}$ on Monte Carlo data set $m = 6$ for $\boldsymbol{\rho} = 0.9$ and $\boldsymbol{\rho} = 0.5$

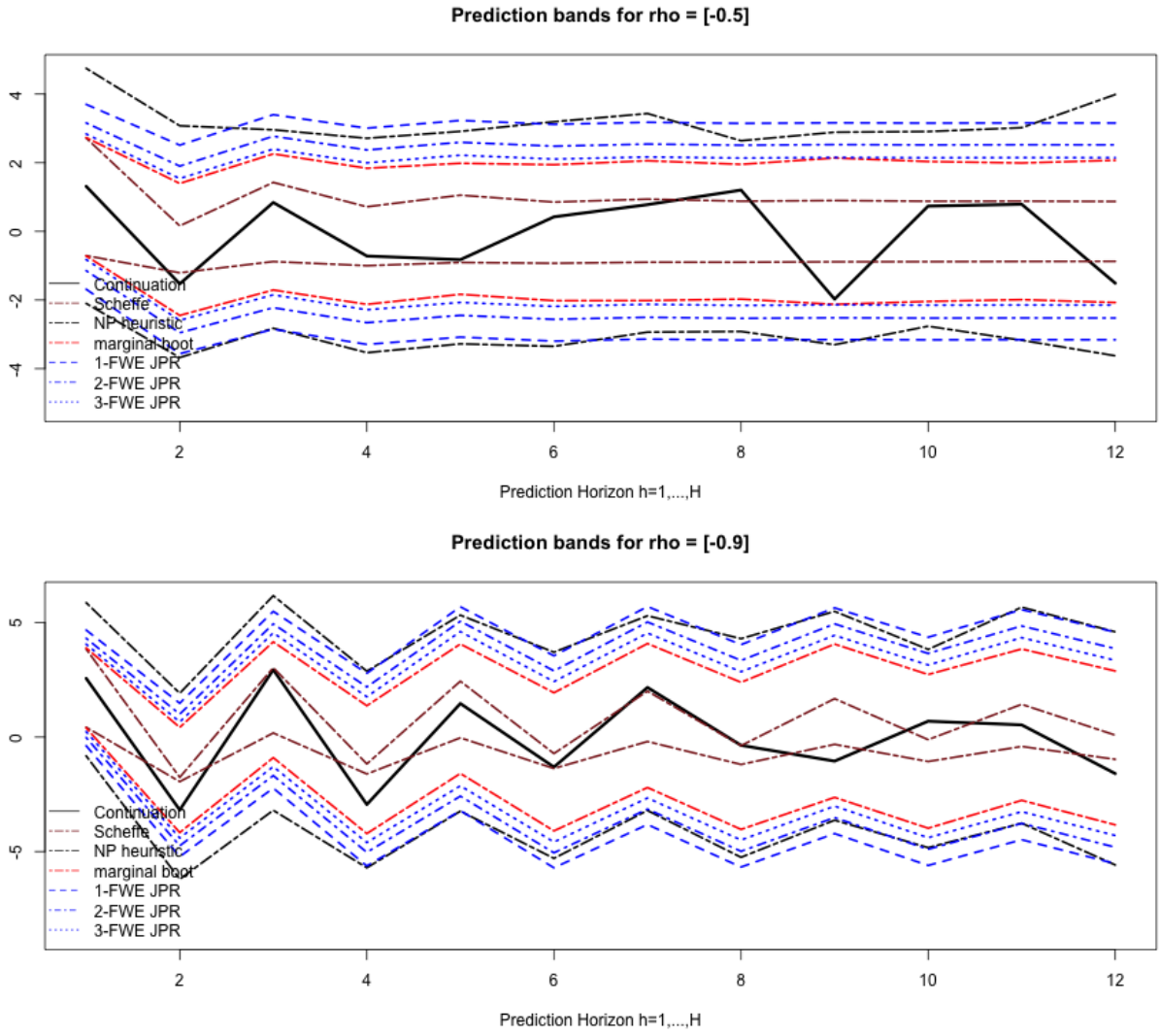


Figure 4.3: AR(1) forecast bands $FCB(\widehat{y}^{(m)}_{H=12})$ and first continuation $y^{(m),(c=1)}_{(T+1):(T+12)}$ on Monte Carlo data set $m = 6$ for $\boldsymbol{\rho} = -0.5$ and $\boldsymbol{\rho} = -0.9$

Part III

Bibliography of All Essays

Bibliography

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.

Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60:953–966.

Beran, R. (1984). Bootstrap methods in statistics. *Jahresberichte des Deutschen Mathematischen Vereins*, 86:14–30.

Bhatia, R. (2007). *Positive Definite Matrices*. Princeton University Press.

Boone, J. and van Ours, J. C. (2006). Modelling financial incentives to get unemployed back to work. *Journal of Institutional and Theoretical Economics*, 162(2):227–252.

Bowden, D. C. (1970). Simultaneous confidence bands for linear regression models. *Journal of the American Statistical Association*, 65(329):413–421.

Bühlmann, P. (2002). Bootstrap for time series. *Statistical Science*, 17:52–72.

Clements, M. P. and Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, 17:247–267.

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22:443–473.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies*, 22:1915–1953.

Gregoriou, G. N., Hübner, G., Papageorgiou, N., and Rouah, F. (2006). Simple hedge fund strategies as an alternative to funds of funds: evidence from large-cap funds. In Gregoriou, G. N., editor, *Funds of Hedge Funds*, Quantitative Finance Series, pages 117–131. Elsevier.

Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management*. McGraw-Hill, New York, second edition.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., and Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the highscope perry preschool program. *Quantitative Economics*, 1(1):1–46.

Heckman, J. J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52:271–320.

Jobson, J. D. and Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36:889–908.

Joehri, S. and Leippold, M. (2006). Quantitative hedge fund selection for funds of funds. In Gregoriou, G. N., editor, *Funds of Hedge Funds*, Quantitative Finance Series, pages 433–454. Elsevier.

Jordà, Ò. (2009). Simultaneous confidence regions for impulse responses. *The Review of Economics and Statistics*, 91(3):629–647.

Jordà, O., Knüppel, M., and Marcellino, M. G. (2010). Empirical simultaneous confidence regions for path-forecasts. Discussion Paper No. DP7797, CEPR. Available at SSRN: `http://ssrn.com/abstract=1611493`.

Jordà, Ò. and Marcellino, M. (2010). Path forecast evaluation. *Journal of Applied Econometrics*, 25:635–662.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, 80:218–230.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Kosowski, R., Naik, N., and Teo, M. (2007). Do hedge funds deliver alpha? a Bayesian and bootstrap analysis. *Journal of Financial Economics*, 84:229–264.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Lalive, R., van Ours, J. C., and Zweimueller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, 3(6):1386–1417.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15:850–859.

Lo, A. (2002). The statistics of Sharpe ratios. *Financial Analyst Journal*, 58:36–42.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Memmel, C. (2003). Performance hypothesis testing with the Sharpe Ratio. *Finance Letters*, 1:21–23.

Pascual, L., Romo, J., and Ruiz, E. (2001). Effects of parameter estimation on prediction densities: a bootstrap approach. *International Journal of Forecasting*, 17:83–103.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–230.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics*, 35(4):1378–1408.

Scheffé, H. (1953). A method for judging all constrasts in the analysis of variance. *Biometrika*, 40:87–104.

Scheffé, H. (1959). *The Analysis of Variance*. John Wiley, New York.

Staszewska-Bystrova, A. (2010). Bootstrap prediction bands for forecast paths from vector autoregressive models. *Journal of Forecasting*.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

Thombs, L. and Schucany, W. R. (1990). Bootstrap prediciton intervals for autoregression. *Journal of the American Statistical Association*, 85(410):486–492.

White, H. L. (2001). *Asymptotic Theory for Econometricians*. Academic Press, New York, revised edition.

White, J. (1961). Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika*, 48:85–95.

Curriculum Vitae of Dan C. Wunderli

Born on 5 December 1982 in Zurich, Switzerland.

1 Education

2008 - Apr 2012	Doctoral Studies, University of Zurich.
July 2010	Institute on Computational Economics, University of Chicago.
2008	lic.oec.publ. (equivalent to two years Master's degree) with distinction (magna cum laude) in Quantitative Finance, University of Zurich.
Summer 2007	Summer School in Advanced Econometrics, London School of Economics and Political Science.
2006	Grundstudium (equivalent to two years Bachelor's degree) in Economics, University of Zurich.

2 Employment

2012 - present	Swiss National Bank
2008 - 2012	Scientific Assistant, Chair for Econometrics and Applied Statistics, University of Zurich.
2006	Internship at the Swiss National Bank, Statistics Department, Current Account Team.